Solid State NMR and Molecular Dynamics studies of solid crystal systems

GLOSSOP, WILLIAM,NELSON

DURHAM UNIVERSITY

DOCTORAL THESIS

Solid State NMR and Molecular Dynamics studies of solid crystal systems

---

Author:
William N GLOSSOP

Supervisor:
Prof. Paul HODGKINSON

A thesis submitted in fulfillment of the requirements

for the degree of Doctor of Chemistry

in the

Department of Chemistry

June 24, 2020

Declaration of Authorship

I, William N GLOSSOP, declare that this thesis titled, "Solid State NMR and Molecular Dynamics studies of solid crystal systems" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

_____

Date:

_____

Abstract

Solid state systems can often exhibit dynamics on the molecular level, changing conformations, rotating or moving across a crystal structure. In this work, a combination of Nuclear Magnetic Resonance (NMR), Molecular Dynamics (MD) and Markov models are utilised to describe and explain the dynamics seen on solid crystals and pharmaceutical solvates. NMR is used first to gain an idea of the dynamics, with relaxation rates used to measure activation energies and motional models suggested to fit the data. MD simulations are performed to directly simulate the motion, with the extraction of copies of each molecule from each simulation allowing motion on timescales greater than the simulation time to be observed. The simulations are analysed utilising Markov modelling, assigning conformations to states, extracting state samples to determine the metastable conformations and calculating transition times between these states.

Several of the lower diamondoids are analysed, providing descriptions of the motion and agreeing with previously suggested models of the dynamics. These also show some of the technical considerations that need to be taken into account when analysing systems using relatively short simulations. However, good agreement with experiment can still be achieved with these methods with careful planning. Two pharmaceutical solvates are also investigated, allowing a picture of the rapid solvent motion to be obtained. These pictures give some interesting avenues for further investigation, as well as showing this method can be applied to different solid systems.

Acknowledgements

I would like to thank my supervisor, Paul, for all his help, advice and guidance during my studies, his suggestions of papers to look into and the discussions of research and other topics. His belief in my presentation skills was a huge boost in Slovenia, and his nudges have helped keep me going to finish this thesis. I would also like to thank Mark Wilson for his discussions on molecular dynamics and his insights into my simulation issues. I would like to thank David Apperley for his teaching of the practicalities of NMR, and for our conversations during a weekly nitrogen fill. I would also like to thank all the members of the SSNMR group at Durham, both under Paul and Karen, for their companionship, advice, and baking skills.

I would like to thank my friends at Durham, in particular the Ustinov College Gaming Society for giving me a weekly relaxation time to unwind from work and far too many snacks on a sunday. My sanity thanks you, although I'm sure I lost that a few times during Eldritch Horror! I want to thank my family for all their love, support, and constant questions about the process of completing a PhD. By explaining it to them, I kept realising just where along the track I was, which helped keep things in perspective and give me goals to work towards.

I would like to thank Holly for her love, support and encouragement. We helped each other complete our degrees, and she has been a constant beacon of light, pushing me to keep going and going until I get this finished. I cannot wait to see what the future holds for us.

I would like to thank my Grandmother, Valerie, my Nan, Francis, and my dog, Leon. I know you would all be proud of me, and I thank you for everything you've done for me. You are always in my heart.

Finally, I thank you, the reader. I hope you enjoy this thesis, and that the work presented here is useful to you in any way it can be.

Contents

Declaration of Authorship iii

Abstract v

Acknowledgements vii

1 Introduction: solid-state motion 1

1.1 Introduction . . . 1

1.1.1 Pharmaceuticals . . . 6

1.1.2 Molecular Machines . . . 9

1.1.3 Markov models . . . 13

1.2 Summary of the current work . . . 18

2 Methodology 21

2.1 Introduction . . . 21

2.2 NMR Details/Methodology . . . 22

2.2.1 Quantum numbers, Angular Momentum and the Zeeman Interaction . . . 22

2.3 MD Details/Methodology . . . 28

2.3.1 Atomistic Simulation Details . . . 30

2.3.2 Force Fields . . . 30

2.3.3 Integration Algorithms . . . 33

Verlet Algorithm . . . 33

Leap-Frog Algorithm . . . 34

Velocity Verlet Algorithm 35
2.3.4 Thermodynamic ensembles 35
Temperature coupling 36
Pressure coupling 37
2.3.5 Equilibration 38
2.3.6 General Simulation Details 38
2.4 Principal Component Analysis 40
2.4.1 Time-lagged Independent Component Analysis 41
2.5 MSM Theory and Application 42
2.5.1 K-Means Clustering 45
2.5.2 Markov Model creation 46
2.5.3 Perron Cluster Cluster Analysis 47
2.5.4 Hidden Markov Models 49
2.6 Use of Python and the PyEMMA module 50
2.6.1 Implied Timescales 51
2.6.2 Transition Path Theory 53
2.7 Interpretation of Results 54
2.7.1 Free Energy Diagrams 55
2.7.2 Metastable partitioning 56
2.7.3 States and state transtions 56
2.7.4 Correlation of Motion 56

3 Diamondoids **61**
3.1 Introduction 61
3.2 NMR Results 66
3.2.1 Diamantane 66
3.2.2 Triamantane 68
3.2.3 1(2)3-Tetramantane 69
3.3 Method Development 70

3.3.1 Diamantane . . . 70
3.3.2 Triamantane . . . 80
3.3.3 Tetramantane . . . 87
3.4 Results . . . 89
3.4.1 Diamantane . . . 89
300 K Simulations . . . 90
Higher Temperature Simulations . . . 92
3.4.2 Triamantane . . . 96
3.4.3 Tetramantane . . . 100
3.5 Conclusion . . . 106

4 Furosemide-Picolinamide **111**
4.1 Introduction . . . 111
4.2 NMR Results . . . 112
4.3 Particulars of Method . . . 113
4.4 FSPA-ethanol . . . 118
4.5 FSPA-acetone . . . 126
^{2}H NMR . . . 136
4.6 Conclusion . . . 136

5 Conclusions and Future work **139**
5.1 Major conclusions . . . 139
5.2 Future work . . . 140

Bibliography **143**

A Appendix A: Data Tables **155**
A.1 Chapter 3: Diamondoids . . . 155
A.2 Chapter 4: FSPA . . . 157

List of Figures

1.1 Comparison of NMR and MD predicted conformations of lorlatinib 5
1.2 Comparison of NMR and MD predicted order parameters of a phospholipid bilayer 7
1.3 Evidence of correlated rotation within a molecular gyroscope 11

2.1 Typical NMR Spectrum 25
2.2 Example date showing directions of PCA components 40
2.3 Reorientation of data using PCA components 42
2.4 Visual comparison between PCA and TICA analysis 43
2.5 Markov Matrix 44
2.6 Illustration of the PCCA method 48
2.7 Example of a Hidden Markov Model 50
2.8 Example implied timescale plot 58
2.9 Example free energy diagram 59
2.10 Example timescale ratio plot 60

3.1 The lower Diamondoids 62
3.2 Triamantane 64
3.3 Tetramantane isomers 64
3.4 Diamantane ^{13}C Spectra 66
3.5 Labelled diamantane structure 67
3.6 Diamantane relaxation data 68
3.7 Triamantane CP Spectra 69

3.8 Tetramantane relaxation data 70
3.9 Tetramantane CP Spectra 71
3.10 Atoms selected for diamantane analysis 73
3.11 Diamantane Positional PCA results 74
3.12 Diamantane analysis vectors 76
3.13 Clustering error example 77
3.14 Diamantane cluster test 78
3.15 Triamantane temperature coupling graph 81
3.16 Triamantane pressure coupling graph 82
3.17 Triamantane Vectors 84
3.18 Triamantane TICA lag time test 85
3.19 Triamantane cluster Test 85
3.20 Triamantane ITS Plot 86
3.21 Tetramantane Vectors 87
3.22 Tetramantane vector plot 88
3.23 Tetramantane ITS Plot 88
3.24 Diamantane raw vector input free energy diagrams 89
3.25 Diamantane principal components - 300 K 90
3.26 Diamantane timescale ratio 93
3.27 Diamantane principal components for all positions - 300 K . . 94
3.28 Diamantane state map - 300 K 95
3.29 Diamantane principal components - 310 K 96
3.30 Diamantane principal components - 325 K 97
3.31 Diamantane principal components - 350 K 98
3.32 Diamantane principal components within the same eigensp-
cae - 300 K and 350 K 98
3.33 Triamantane TIC plot for position A 99
3.34 Triamantane FED plots 99
3.35 Triamantane state map - 400 K 101

3.36 Tetramantane TIC plot - 350 K 102
3.37 Tetramantane principal components - 350 K 103
3.38 Tetramantane state map - 350 K 104
3.39 Tetramantane example etates - 350 K 105

4.1 FSPA crystal structure 113
4.2 Fitted FSPA-Ethanol relaxation data 114
4.3 FSPA-Acetone Lineshape Analysis 115
4.4 Fitted FSPA-Acetone relaxation data 116
4.5 FSPA-Acetone analysis vectors 118
4.6 FSPA-Ethanol analysis vectors 118
4.7 FSPA-Ethanol TICA lag time test 119
4.8 FSPA-Ethanol cluster test 120
4.9 FSPA-Ethanol TIC plot - 273 K 121
4.10 FSPA-Ethanol ITS plot 122
4.11 FSPA-Ethanol state map 124
4.12 FSPA-Ethanol state samples 124
4.13 FSPA-Acetone TICA lag time test 126
4.14 FSPA-Acetone cluster test 127
4.15 FSPA-Acetone TIC plot 1 - 273 K 128
4.16 FSPA-Acetone TIC plot 2 - 273 K 129
4.17 FSPA-Acetone ITS plot 130
4.18 FSPA-Acetone timescale ratios 131
4.19 FSPA-Acetone state map 131
4.20 FSPA-Acetone state map 2 132
4.21 FSPA-Acetone state samples 133
4.22 FSPA-Acetone state map - 323 K 134

A.1 Triamantane TIC plot all positions 156
A.2 FSPA-Ethanol Timescales ratio plot 157

List of Tables

3.1 Diamantane mean first passage times - 300 K 91
3.2 Triamantane mean first passage times - 400 K 101

4.1 FSPA-Ethanol mean first passage times - 273 K 125
4.2 FSPA-Ethanol mean first passage times - Variable temperature 125
4.3 FSPA-Acetone mean first passage times - 273 K 130
4.4 FSPA-Acetone mean first passage times - variable temperature 134

A.1 Diamantane mean first passage times - 310 K 155
A.2 Diamantane mean first passage times - 325 K 155
A.3 Diamantane mean first passage times - 350 K 155
A.4 Diamantane correlation coefficient data - 350 K 156
A.5 Triamantane full transition time table 156

List of Abbreviations

NMR	**N**uclear **M**agnetic **R**esonance
MD	**M**olecular **D**ynamics
PCA	**P**rincipal **C**omponent **A**nalysis
TICA	**T**ime **I**ndepedent **C**omponent **A**nalysis
FSPA	Furo**S**emide **P**icolin**A**mide
FED	**F**ree **E**nergy **D**iagram
ITS	**I**mplied **T**ime**S**cales
CP	**C**ross **P**olarisation
QM	**Q**uantum **M**echanical
HMM	**H**idden **M**arkov **M**odel
MSM	**M**arkov **S**tate **M**odel
SSNMR	**S**olid **S**tate **N**uclear **M**agnetic **R**esonance
FID	**F**ree **I**nduction **D**ecay
CSA	**C**hemical **S**hift **A**nisotropy
PAS	**P**rincipal **A**xis **S**ystem
RF	**R**adio **F**requency
EFG	**E**lectric **F**ield **G**radient
PCCA	**P**erron **C**luster **C**luster **A**nalysis
TPT	**T**ransition **P**ath **T**heory
TIC	**T**ime **I**ndependent **C**omponent
PC	**P**rincipal **C**omponent
XRD	**X**-**R**ay **D**iffraction

Dedicated to everyone who supported me.

1 Introduction: solid-state motion

1.1 Introduction

Solid materials appear to be immobile at the macroscopic scale. However, at the atomic level, there is motion, with the lower limit of this in terms of energy being lattice vibrations. Most molecules in solid materials exhibit more dynamics than this, with bond angles and lengths stretching, compounds altering their orientations or conformations, and even moving across a crystal structure. These motions can be crucial processes in the material, and can give rise to many of the useful (or not so useful) properties of the solid.

There are several tools available to the chemist who wishes to study dynamics in materials. Nuclear Magnetic Resonance (NMR) has long been used to study solid materials and the dynamics within them, being particularly suited for this task. Various experiments allow investigation into slow motions (Hz timescale), fast motions (MHz-GHz timescale), or a combination of both, as they can both occur within solid-state materials. Additionally, as NMR records a collective signal from the entire sample, it records the local environment of every distinct atom type present in the sample is recorded, giving us access to all this information. Samples can also be labelled isotopically, and the techniques tuned to particular nuclides, so structure-specific information can be obtained about the a portion of the sample at an atomic level.

NMR does have its drawbacks however. The results obtained from the experiments do not directly link to the precise motions involved, instead requiring us to postulate particular motions and then fit these results to models. These models can be drawn from other experimental techniques, or simply by considering the molecules involved and determining the most likely motions available to them. Other techniques, such as Raman spectroscopy, share these problems, allowing information to be gathered about a particular system, but being unable to link this to exact molecular motion.

The complexity of the system in question also determines our ability to construct models. Simple molecules and systems often have intuitive motions available to them, which we can use to create models that allow us to obtain parameters relevant to the experimental results. For example, methyl groups undergo a C_3 rotation, something that is chemically intuitive from their shape, and so we can use a 3-site jump model to explain our results. However, as the system becomes more complex, determining suitable models of motion becomes increasingly tricky. As such, we can apply computational techniques to the task, namely Molecular Dynamics (MD) simulations. These simulations will model the dynamics present within complex systems, allowing us to observe motions from the picosecond range up to the hundreds of nanoseconds. MD can directly "see" the motions of the atoms and molecules, providing us with models to fit the experimental results to, and allowing us to study these processes directly with a level of detail that experiments would struggle to match. While some questions can be raised about the simulation quality, comparison of the results with experimental data can validate the simulations and give credibility to the analysis performed upon them.

The data produced by MD simulations consists of a series of coordinates of

atoms arranged in chronological order as a trajectory. Transforming this trajectory into useful information requires in-depth analysis of the motions described by the trajectory and the timescales present within them. For this purpose, the construction and usage of Markov models has been proven to be both useful and effective [1]. Bowman *et al.* applied Markov models to the study of proteins, by simulating the motion of TEM-1 β-lactamase using MD techniques [2]. The resulting trajectory was used as the basis for a Markov model to describe the possible conformations the protein could adopt. To do this, they used the combined data from a thousand simulations, to give an aggregated simulation time of 81 μs. This allowed them to achieve enough sampling of the conformational space to obtain a reasonable description of the likely conformations available, and utilised multiple simulations or replicas, rather than a single long trajectory, a key point. The next step was clustering, the assignment of each conformation to a particular state, and to speed up this process, every 10th frame of the simulation was used, with the rest of the data assigned to the clusters generated. The clusters can be analysed by querying the model to identify local fluctuations in the conformations that produce sites that mimic the active site. By allowing the system to explore the conformational space to a large degree, and then going through the process of clustering and extracting the states corresponding to each cluster, they were able to discover the presence of these hidden allosteric sites. Hidden allosteric sites are difficult to discover experimentally, and so the combination of MD and Markov model techniques allows insight that would be lacking otherwise.

Zhuang *et al.* have utilised Markov models in conjunction with MD simulations to recreate and predict experimental data, specifically the result of a T-jump experiment. A T-jump experiment involves using an intense laser pulse to rapidly heat a sample, which can be used to trigger a process that is

driven by a large enthalpy change. Zhuang constructed two sets of Markov models, one each for before and after the T-jump, and then matched states between them. Assuming the temperature change is very fast, there should be no noticable difference between the state populations before and after the T-jump. Therefore, they used the distribution among states from the low temperature model, and applied the transition probabilities of the high temperature model to this, to simulate the effect of rapidly changing the temperature. The changing conformation of the protein after the jump was then used as the basis for predicting IR and 2DIR spectra, which matched nicely when compared with experimental spectra [3].

Markov models partition the data into discrete states, counting transitions between them and using these to predict the probability of moving from one state to any other state within a given time frame. This data can then be analysed further to calculate transition times between any specified pair of states, as well as the most likely path from state to state that the transition could take [4, 5]. Clearly, applying these models to the problem of describing the motions would result in almost direct extraction of the data of interest. Independently, Markov models are used to simulate NMR spectra of systems (particularly those containing ^{2}H) affected by dynamics. One implementation is the EXPRESS program, which utilises the idea of Markovian jumps to simulate experimental spectra [6]. However, such programs typically require the user to manually enter in the molecular orientations and jump rates, and these will need to be determined either via a different experiment, or using a different technique.

NMR and MD are highly complementary techniques, each technique either validating or explaining information gained by the other technique, and they have been combined together successfully many times. Biological membranes have been studied using this combination [7], as dynamic behvaiour

explains the majority of the functionality of such systems. Simulating particular biological systems in order to extract their NMR parameters based on their dynamics has yielded results in excellent agreement with experiment [8]. This technique, using MD simulations to predict useful NMR data, has been applied to systems as diverse as macrocyclic drugs and lipid bilayers with great success [9, 10]. To demonstrate this, I will detail two examples.

Peng investigated the conformations of the lorlatinib macrocyclic drug, commonly used to treat nonsmall cell lung cancer. To do this, they utilised solution-state NMR experiments with replica exchange MD, a technique that enhances the sampling of conformational space by exchanging neighbouring replicas of the simulated molecule at different temperatures. NMR experiments were used to calculate interproton distances through the use of nuclear Overhauser enhancement experiments, whereby nuclear spin polarisation is transferred between nearby spin-active nuclei. These distances can then be used to predict possible conformations, and compared to the conformations produced by the MD simulations, resulting in a high level of agreement in both polar and non-polar solvents (see figure 1.1) [11].

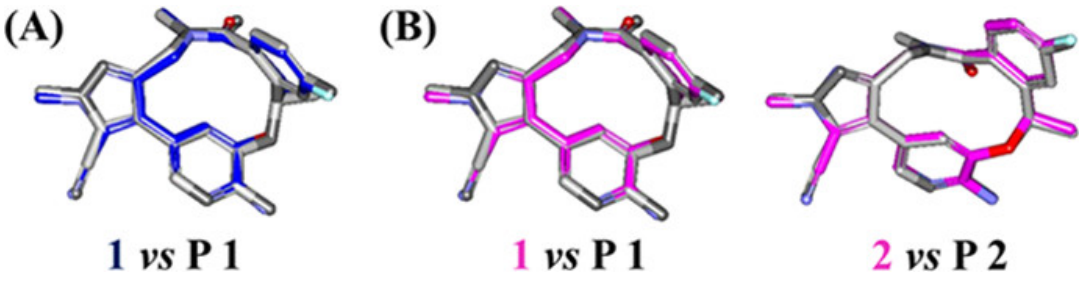

FIGURE 1.1: Solution conformations of lorlatinib in D_2O/DMSO-d6 (blue) and $CDCl_3$ (pink) aligned with the predicted conformations (gray). The molecule was divided into two parts, P1 and P2 for the MD analysis, to determine the conformation most suited to describe each part. The NMR analysis predicted 1 conformer in D_2O/DMSO-d6 (**1**) and 2 conformers (**1** and **2**) in $CDCl_3$. The figures show: A) **1** vs P1 in aqueous solution, (B) **1** vs P1 (left), and **2** vs P2 (right) in chloroform. Figure obtained from reference [11]

Vermeer investigated phospholipid bilayers using the combination of NMR and MD techniques. Phospholipid bilayers are a key component of biomolecular membranes, and so understanding them and their interactions can be vital to predictions of drug interactions or bioavailability. NMR was used to calculate order parameters, a measure of the dynamics of the lipids, and these parameters can be calculated directly from the quadrupolar splittings of a ^{2}H solid state NMR experiment. MD simulations can be used to calculate these values too, by directly measuring the angle between a specified C-D bond vector and a given reference axis. These values can be compared, in order to draw conclusions about the accuracy of the MD simulations, and whether further predictions from the MD can be trusted. Figure 1.2 shows the comparisons of these values for three different systems. As can be seen, there is a high level of agreement between the values, and while some values do not match entirely, an important point is that the trends in the values are followed, allowing general conclusions to be drawn, even if specific numbers cannot be produced [12].

1.1.1 Pharmaceuticals

Research into pharmaceuticals is a key area of interest, as we rely on the production and behaviour of drugs for significant portions of modern medicine. The field of NMR crystallography is concerned with using NMR to shed light on crystal structures, and is especially able to resolve disorder within structures. Disorder in crystals is fairly common, and can be both static (where the disordered fragments remain in distinct orientations), or dynamic (where the disordered components either continously move or jump between particular orientations). Some recent applications of using NMR to study dynamics and disorder in pharmaceutical solids include the use of ^{13}C T_1 relaxation times to characterise ring disorder [13], while measuring the change of relaxation

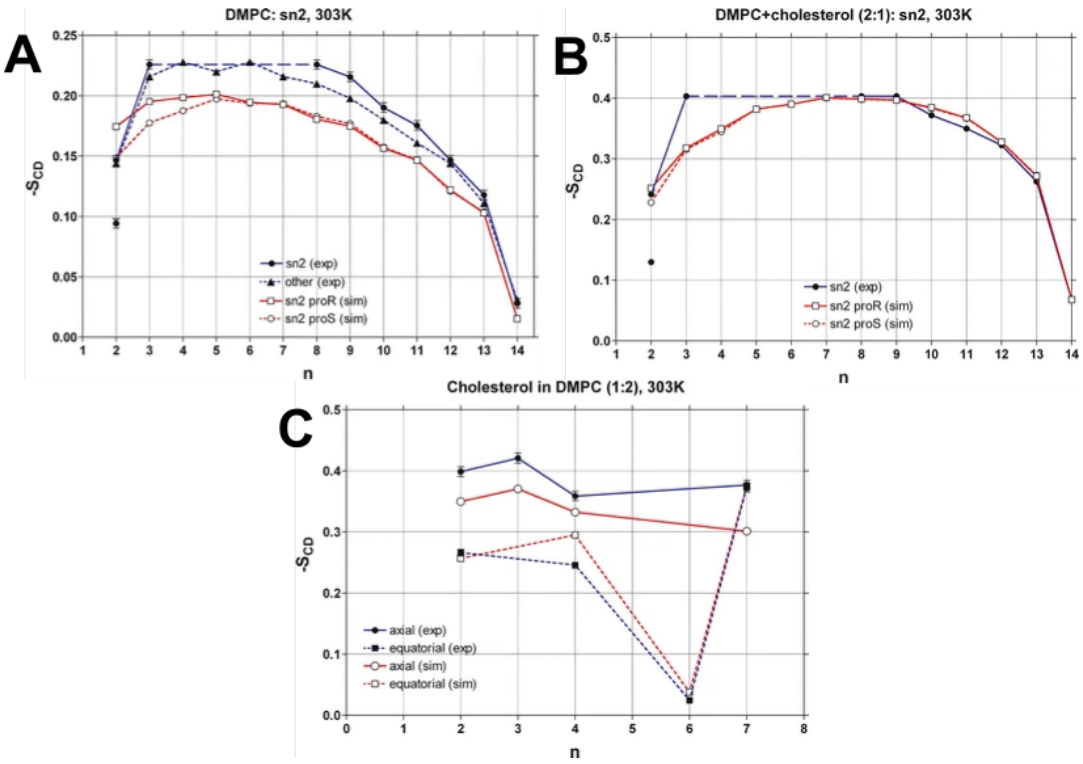

FIGURE 1.2: Graphs comparing the experimental (NMR) and simulated (MD) order parameters for the DMP phospholipid bilayer. A refers to the order parameter -S_{CD} for the carbon chain of pure DMPC at 303 K as a function of the carbon atom index. B refers to the order parameter -S_{CD} for DMPC in the DMPC-cholesterol 2:1 mixture at 303 K as a function of the carbon atom index. C refers to the order parameters -S_{CD} for cholesterol in the DMPC-cholesterol 2:1 mixture as a function of the carbon atom index. Figure adapted from reference [12]

times with temperature and fitting this to an Arrhenius curve has allowed the calculation of activation energies and conclusions about the most likely form of molecular mobility to be drawn [14]. When crystal structures contain solvent molecules, the resulting structures are termed solvates and the solvent molecules themselves can show interesting dynamics. The pharmaceutical droperidol has a range of solvates, and the use of ^{13}C and ^{2}H spectra and relaxation times gave various insights into the dynamics [15]. The measurement of relaxation times give an indication that the motion is on the order of 10s MHz, while the minimal averaging of the ^{2}H quadrupolar parameters

indicate the motion has limited amplitude. Additional dynamics were observed using the ^{13}C spectra, with these dynamics on the order of 10s kHz, and most likely to be flipping to an equivalent state through inversion symmetry. This shows NMR's ability to see multiple timescales of dynamics. The use of NMR alone allows for the measurement of dynamic processes, and combined with selective labelling, can readily provide activation parameters and timescales [16]. Subsequent fitting of the data to models can be trickier, as shown in the paper. The phenyl ring motion is fairly easy to predict, undergoing a two-site jump, while the motion of the fumerate is harder to model, as the motion is could exhibit is not obvious. Even without selective labelling, the tools available to the spectroscopist allow the dynamics at multiple solvate sites to be observed, with subsequent fitting suggesting available motions [17].

Molecular dynamics has also been used extensively within the pharmaceutical industry, in both drug design, analysis and characterisation [18–20]. While initial forays into combining MD with biological systems yielded short timescales [21], the increase in computational power has allowed significantly larger simulations, increasing from several hundred atoms in the late 70s to 50,000-100,000 atoms being routine, and even 500,000 atoms with the right computational hardware [7]. Simulations have investigated the interactions between drugs and biological membranes, particularly binding interactions, probing their energy values and the processes by which they bind [19, 22, 23]. Polymorphism within pharmaceuticals is of particular interest, and can be studied through the application of MD techniques [24]. Often these techniques start with a simulation of a disordered pharmaceutical, and the temperature during the simulation is lowered to facilitate crystallisation. These glasses are then simulated further, with bonding patterns, bonding

energy and diffusion coefficients able to be extracted from the resulting simulations [25]. The effect of different solvents directing the crystal form of 5-fluorouracil was investigated, with hydrogen bonding with water promoting the formation of form I, while dry nitromethane facilitates the structure of form II [26].

1.1.2 Molecular Machines

Artificial molecular machines have gained considerable attention for some years. With an aim of emulating nature by using collections of interacting molecules to produce supramolecular functionalty, the field has steadily grown [27–29]. As a comparison to regular machines, these systems would be designed to react to specific external stimuli, such as changes in local magnetic fields or light conditions, to produce a useful response. Chemists have produced considerable work on this topic, synthesising molecules that display properties similar to the components of larger scale machines. There have been examples of molecular gears [30], turnstiles [31], gyroscopes [32] and even molecular motors [33]. These machines can be combined and designed to have applications as diverse as switchable dielectrics, gas seperation and chemical sensing [34]. However, in order to control the properties of the machines made up of the individual components, accurately characterising the kinetic behaviour is vital.

In these cases, the form of the motion is well-defined by the design of the components themselves, and so NMR is used to extract the kinetic parameters of the motions with as high an accuracy as possible. NMR is perfect for this task, and is a central technique used to extract these parameters [35], due to the wide range of NMR epxeriments available to study the rate of these dynamics on different timescales. More specifically, studies of potential components for various molecular machines have utilised: ^{13}C and ^{2}H T_1

relaxation measurements to extract the energetic barriers to rotations occurring at the rates of MHz [36, 37], ^{2}H line shape analysis to obtain information about dynamics with rates of 100s of kHz [38], using models to predict the orientations that contribute to observed dynamics [37–39], and using the coalescence of peaks in ^{13}C spectra can be used to obtain information of far slower rates, in the realm of 100s of Hz [36, 38–40].

As the motion is designed and predictable in these systems, MD simulations are often used to gather more information about potential energy landscapes of the components as a function of degrees of freedom. There are various methodologies available for this, although the general technique is to identify a degree of freedom that is of interest, and then calculate the energy of the molecule as this degree changes over the course of an MD simulation. In the case of *Zimmerman*, intermediate reaction structures were generated using MD methods, and the structures then extracted and underwent a restricted structural energy minimisation. The energy of interaction between these structures and the general lattice was calculated, and found to be in very good agreement with the stereochemistry of the reaction [41]. In the case of *Jarowski*, they identified the dihedral angle as the reaction co-ordinate of interest, and then used constrained MD simulations (simulations during which a particular physical property e.g. bond length, bond angle, dihedral angles etc. are fixed to a particular value) to scan through the entire range of dihedral angles, taking energy values after every step. This technique allowed them to compare experimental and theoretical energy barriers, with experimental energies of 12.8-14.6 kcal/mol recorded and theoretical barriers of 15.5-16.2 kcal/mol calculated. They then went further, studying correlated motions within the system by following the motion of the atoms within the model, except for the central atoms which the simulation is fixing in place. As shown in figure 1.3, there was a clear increase in root-mean-square distance

and maximum displacement as the central atoms were rotated, followed by a decrease as they approached a rotation angle of 180 °. This showed that as the central atoms rotated, the surrounding structure moves to accomodate it, before snapping back into place again [42].

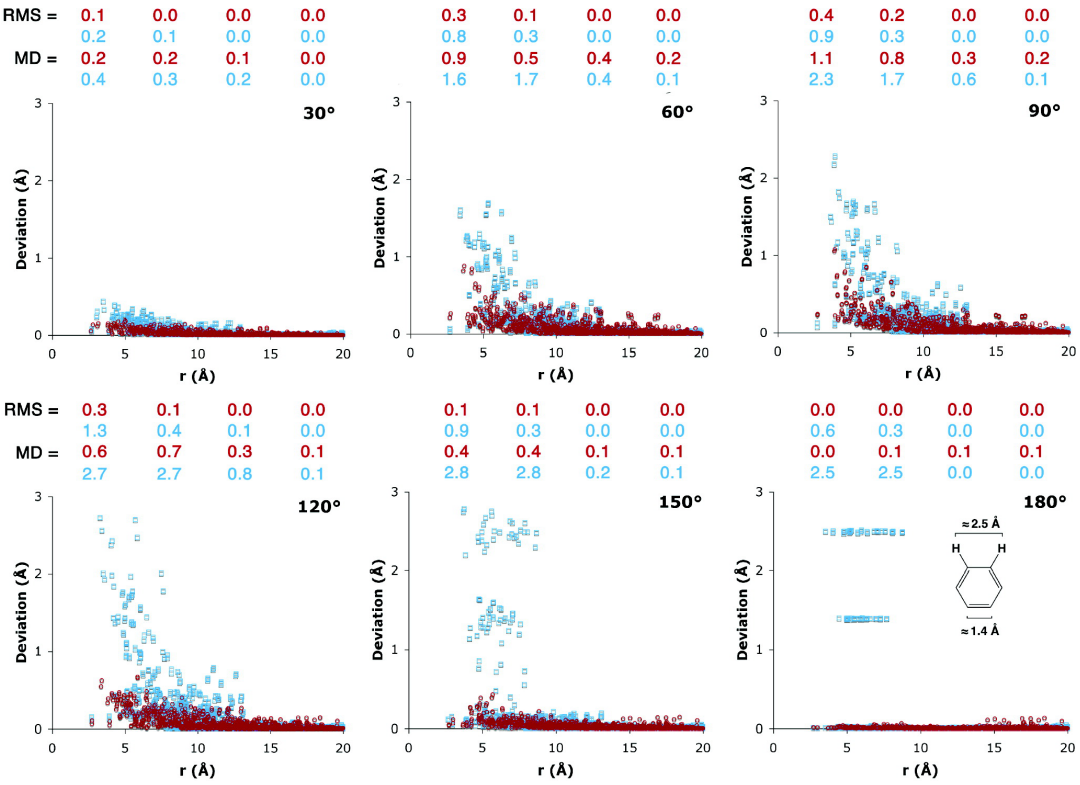

FIGURE 1.3: The root-mean-squared (RMS) and maximum displacement (MD) values for all atoms surrounding a central rotating ring of a crystalline molecular gyroscope, either with benzene as a solvent (blue numbers and points) or in a desolvated state (red numbers and points). The increase and decrease of values as the ring rotates shows the motion of the surrounding atoms are correlated to the rotation of the rotor. Figure obtained from reference [42]

Combining NMR and MD can allow us to extract a whole host of data, and combining these techniques with others, such as X-ray diffraction, has facilitated the study of octafluoronaphthalene, with remarkable success. The combination of NMR and MD simulations allowed the dynamics to be described, with the NMR observing two distinct processes of different timescales, while the MD describes the nature of the motion. This study also shows that long

timescale motions, several orders of magnitude larger than the simulation length, can be accessed via classical molecular dynamics, through the use of combining the individual trajectories of every molecule within the simulated system to achieve a total trajectory time of 14.4 μs from a single 100 ns simulation [43]. The study was continued using enhanced sampling methods, namely metadynamics, to form a picture of the free energy surface. Metadynamics enhances the exploration of the free energy space by biasing the simulation against previously explored areas, forcing the system to explore further. By performing a metadynamics simulation, and then inverting the resulting bias potential map, the full free energy surface can be produced, more than 4 times quicker than classical MD simulations [44]. Additionally, by expanding the criteria against which the system is biased, the effect of correlations between molecules can be investigated. All this is achieved in simulations of fewer than 50 ps, compared to an unbiased simulation of 100 ns that failed to replicate the results. This is just one example of applying further analysis techniques to MD simulations to enhance the quality and descriptiveness of the data obtained.

MD simulations have been used to examine more complicated forms of motion, providing appropriate models of the motion and supporting the results gained from the NMR studies [37]. Diffusion within host-guest structures can be observed through MD, with simulations allowing us access to detailed descriptions of the systems. Extracting coordinates allows diffusion coefficients, order parameters and torsion angles to be obtained, which can be combined with time-sensitive data to develop clear pictures of the movement within systems [45]. NMR experiments can be performed to provide specific data, typically motional correlation times, that is used to verify MD simulations by comparing the difference between the experimental and computational results. If these results are in agreement, these simulations can be

used to directly view the motions the molecule is exhibiting, via analytical methods or simply viewing the trajectory. In either case, performing statistical analysis on the trajectories can be the basis of predictions of the dynamics and motions of solid systems [46, 47]. These techniques are often applied to protein models, with the method of principal component analysis (PCA) combined with MD analysis to obtain results. *Leitner* extracted protein dihedral angles from simulation and used these values in conjunction with PCA to create an energy landscape that describes the dynamics of the tetrapeptide IAN [48]. This technique was found to converge quickly after a relatively short simulation, which allows for rapid multiple simulations of the system. These repetitions of simulation can allow us to acquire enough data to construct Markov models. *Altis* utilised the technique to first construct a free energy landscape using the dihedral values in conjunction with PCA, then applied a clustering algorithm and constructed a Markov model transition matrix from the resulting data, allowing a low-dimensional description of the alanine dipeptide motion [49]. *Jain* again used PCA on the dihedral angles of a villin headpiece subdomain, but split the protein into parts and utilised PCA on each part separately. After clustering, a Markov model was produced, with each different combination of each part representing a unique conformation of the entire protein, and reconstructing the folding time of the protein [50].

1.1.3 Markov models

The idea of a Markov chain was originally put forward in 1906 by Andrey Markov [51]. While originally used exclusively in the mathematical sciences, Markov chains are now used in chemistry, biology and physics, finding a number of different uses. One particular use was to model reaction networks as Markov processes, using the number of molecules of each species as states,

and reactions as possible transitions [52]. While these initial forays investigated straightforward kinetics, the use of a Markov chain has been expanded to include more complicated dynamics.

The mathematical theory of estimating Markov models from MD simulations was introduced by *Schuette* [53], combining the theory of Markov models with data from Monte Carlo simulations [54]. In constructing these Markov models, the aim is to achieve a description of how the system explores the conformations available. Ideally, we hope to reach an equilibrium and construct a reversible Markov chain, that is, one that exhibits detailed balance. The principle of detailed balance for a kinetic system, such as those simulated by MD, states that at equilibrium, each elementary process is in equilibrium with its reverse process. This means that the transition matrix produced is essentially symmetric, so that rate of moving from state a to state b is the same as moving from state b to state a. By sampling every possible move from a given point, we can fully explore the configuration space of a given transition matrix [55]. These ideas have been applied to protein structures, where Markov models are used to generate the next step in the folding sequence for a protein [56]. To illustrate these ideas, I will consider the example of *Bratholm* and *Jensen*. In their work, they sought to improve the prediction of protein NMR chemical shifts using quantum mechanics based methods. To do this, they utilised QM based predictions of the chemical shifts from an initial starting structure, then altered the structure using a Markov Chain Monte Carlo method. The method edited the starting structure according to a set probability of structural changes (the Markov chain portion of the method), then accepted or rejected this change based on improvements in the match between experimental and predicted chemical shifts (the Monte Carlo portion of the method). By utilising these methods, the difference between predicted and experimental chemical shift for 17 proteins reduced by

an entire ppm for carbon and nitrogen, and 0.15 ppm for hydrogen [57]. This framework of using the agreement of a suggested structure with experimental data as the acceptance criteria for a move has been used previously [58, 59]. Besides proteins, these methods have been applied to solid state systems as well, with recent work observing the reversal of growth of molecular crystals [60] and solid transition metal complexes[61].

Swope described a method of utilising multiple MD simulations together to examine protein folding kinetics, by extracting properties of interest, assigning them to states and using the relative probabilities of these states to construct a Markov chain. From this, motions between these states and the timescales associated with them have been found, which can be related to the protein's specific function[62, 63]. Other work has expanded on traditional Markov models to hidden Markov models (HMMs), that is, models in which we cannot observe the underlying process directly, but can see the outcomes of the process. For example, a molecule could adopt a range of very similar orientations, but these orientations contribute to a single metastable state. The specific orientation is the underlying or hidden state, while the metastable state is the observable we can see. Once the model is complete, folding times and the transition pathway as the protein folds can be extracted, both pieces of information which are vital in understanding protein dynamics [64]. The transition matrix obtained from generating a Markov model can be analysed further, showing that the system needs to adopt a particular conformation before accessing further ones, dividing the system into active and inactive states, which can affect the kinetics of the protein folding[65].

Since then, the idea of combining MD simulations with Markov models has become common, using the simulations to observe the dynamics and the MSMs to obtain kinetic and thermodynamic data as well as define metastable

states [66]. These states are not limited to single molecules or even copies of the same molecule: binding behaviour between trypsin and benzamide was studied by *Noé* in 2015. A MSM was constructed based on a 149.1 microsecond simulation of the system, which allows calculation of binding/unbinding rates and the free energy of binding [67]. Markov models and MD together can probe molecular reactions by studying the states available to a particular reactive system. By observing or biasing the simulation towards the end result of a given reaction, a list of intermediate conformations can be produced. By observing the transitions from starting structures, to intermediate states, to final structures, an MSM can be constructed from this data to obtain reaction coordinates for further simulation and analysis [68, 69]. For example, say a reaction can be described by $A + B \rightarrow C$. By biasing the simulation towards C, using intermolecular distance for example, we might see that A and B form an intermediate, metastable complex AB before moving on to form C. A Markov model can then be generated from the data, showing the probability and transition time of forming the intermediate, and then moving to the end result C. More general biomolecular dynamics can also be probed, with techniques that streamline the process of clustering data into the states required to construct a Markov model. As clustering is a key step in the model construction process (see section 2.5.1), a general approach that maintains high accuracy is of some worth [70].

The idea of using MD to provide the data necessary to construct a Markov model is key, and has been proven to work well. This is due to MD's ability to allow systems to explore their conformational space either with bias, to ensure the entire space is explored, or without bias to reproduce the thermodynamic properties accurately. By exploring the conformational space and then reproducing the thermodynamics, the states and transitions between them can be calculated, which results in the construction of a Markov model.

[71, 72]. In order to construct an accurate Markov model from MD data, a large number of data points is required. With a large data set, the transition matrix can accurately reproduce the behaviour of the system, whereas with a smaller data set, the system diverges quickly from Markovian behaviour [73]. However, even limited MD data can allow the construction of Markov models that give information on processes an order of magnitude longer than the simulation time.[74–76]. This process involves performing multiple simulations of short time length to construct an accurate Markov model, and then using this model to rapidly simulate the system for longer timescales. Simulating for the time required to see the long process during the simulation is possible, albeit at a far increased time cost, whereas performing multiple short-length simulations can produce the same data [77, 78]. This can also apply to extracting individual trajectories of molecules within one simulation, and using these to produce the Markov model.

While a great deal of research has applied Markov model ideas to the study of proteins, Markov chains have also been applied to dynamics of solids and molecular crystals. In particular, the growth of molecular crystals has been extensively modelled using Markov chains [79]. The formation of crystals can be modelled as a Markov chain process, with the probability of a particular molecular alignment as the crystal grows relying only on the current polarity of the connecting point [80]. This idea was expanded to study aperiodic crystals. Aperiodic crystals cannot be described by a simple unit cell, as while they are symmetric in 2 dimensions, they are irregular in the third. The type of layer was found to be dependent on the previous layer, and as such the sequence of each layer can be transformed into a Markov model. This technique has been applied recently to ice and opaline silica, and the distribution of faults in the crystals and number of layers that influence the current layer can be predicted [81, 82]. The self-assembly of colloidal systems can be

modelled in a similar way, although the construction of an accurate Markov model is required beforehand. To achieve this, MD simulations of colloidal motion can be utilised [83]. MD simulations provide ample data with which to construct a Markov model, and can be used repeatedly to analyse different aspects of the molecule, so one simulation can be used to produce multiple models.

As I have shown, Markov models have been extensively used with biological systems in conjunction with NMR, and has also been applied to various solid systems. In this thesis, I hope to take these ideas and apply them to solid-state dynamics, utilising the techniques to extract useful kinetic and dynamic data about several chemical systems.

1.2 Summary of the current work

Chapter 2 outlines the various techniques used in this work. NMR spectroscopy is introduced and relevant interactions described, as well as describing various experiments that can be used to obtain data on molecular dynamics. The technique of Molecular Dynamics (MD) is also outlined, explaining the basic theory behind it and describing the methodology of the simulations. The chapter ends with sections on Markov State Modelling and the various techniques associated with it to characterise and interpret the data. Chapter 3 presents data taken from NMR experiments and MD simulations on diamantane, triamantane and tetramantane. The combination of NMR, MD and MSMs is applied these system, for which a reasonable amount of information is known. After performing several simulations, the key orientations of the molecules and the transitions between these states is measured, and are in excellent agreement with experimental results. States are extracted and the temperatures at which new motions arise are identified. Chapter 4 presents results from studies on solvates of the furosemide-picololinamide cocrystal.

The method outlined previously required some adaptation to be applicable to the new system, but after applying these changes, states and transition times between states are also extracted. These results are then linked back to NMR studies, as well as signposting further directions to research. Chapters 3 and 4 are also being prepared for publication.

2 Methodology

2.1 Introduction

A key focus of the work undertaken is to combine SSNMR data with various computational methods in order to obtain a greater understanding of the dynamics present in solid systems. Ideally, this method would be simple to use, with minimal training required to utilise and interpret the data, rapid, with a full set of data able to be obtained in a short space of time, and accurate, with the method providing reasonable suggestions and results. Developing a method with all three qualities is no easy feat, and some compromises may have to be made. Speed in data collection often requires training and familiarity with the software, whereas accuracy may require extended periods of data collection. Finding the right balance between these three is therefore essential.

The method developed hopes to strike this balance. It requires some training in computational methods and SSNMR experiments, but the simulations performed are not overly complex. In ideal circumstances, the simulations can take less than a week from construction to completion, although the NMR experiments can have varying lengths, depending on the properties of the system. Finally, if constructed correctly, the simulations can provide accurate representations to go along with the carefully measured spectra, providing a well-rounded view of the motion of the system.

In this chapter, I will go into the theory behind the various underlying techniques, starting with the SSNMR details and moving onto molecular dynamics simulations. An explanation and discussion of the use of the statistical methods of Principal Component Analysis (PCA) and Time-Independent Component Analysis (TICA) will follow, as well as describing the usage of Markov State Models (MSMs) to extract states of interest from the overall trajectory. Finally, the combination of the techniques will be explained, showing how to move from a starting sample of interest and structure file, to a full set of data showcasing the dynamics of the crystal.

2.2 NMR Details/Methodology

Nuclear Magnetic Resonance (NMR) spectroscopy is a widely used chemical characterisation and imaging technique. While most modern chemical laboratories focus on solution-state NMR, largely for determining the products of reactions, there is a growing usage of solid-state NMR to characterise insoluble compounds or to observe dynamic behaviour that would be lost when dissolved in solution. In particular, crystal dynamics cannot be observed in solutions, as the crystal structure is entirely lost, leaving solid-state NMR as the technique of choice for these purposes.

This section gives an overview of the NMR interactions, indicating the differences between solution-state and solid-state techniques and their spectra.

2.2.1 Quantum numbers, Angular Momentum and the Zeeman Interaction

A moving or spinning charge, such as those generated by a nucleus of non-zero magnetic quantum number, produces a magnetic moment. When this is put into a high magnetic field (the direction of which is usually taken to be

the z axis), the moment is separated into $2I + 1$ different states, each defined by the projection quantum number m_I. Using the $I = \frac{1}{2}$ system for simplicity, there are two states: α, where the spin is aligned in one direction with respect to the magnetic field (also labelled as m_I = 1/2) and β, where the spin is aligned in the opposite direction to α (also labelled as m_I = -1/2). The energy of these two states is related by the Zeeman Effect, and can be expressed as:

$$E_z = \frac{-h\gamma m_I B_0}{2\pi} \tag{2.1}$$

Where E_z is equal to the Zeeman energy, h is the Planck constant, γ is the gyromagnetic ratio of the nucleus, m_I is the projection quantum number and B_0 is the strength of the magnetic field.

The difference in energy between the two states in the system is :

$$\Delta E_z = \frac{-h\gamma B_0}{2\pi} \tag{2.2}$$

This can be transformed into an expression for the resonance frequency, ν_{nmr}, using the equation:

$$\Delta E_z = h\nu_{\mathrm{nmr}} \tag{2.3}$$

Where:

$$\nu_{nmr} = -\frac{\gamma B_0}{2\pi} \tag{2.4}$$

This frequency is dependent on the value for γ and the field strength of the external magnetic field, and is known as the Larmor frequency. As a chemical system is rarely made up of a few isolated atoms, a description of the bulk of the sample is required. The distribution between the states will follow a Boltzmann distribution:

$$\frac{P_\beta}{P_\alpha} = e^{-\frac{h\nu_{\mathrm{nmr}}}{k_{\mathrm{B}}T}} \tag{2.5}$$

where P_α and P_β are the populations of the α and β states respectively, k_B is the Boltzmann constant, and T is the temperature. This ratio depends on both the strength of the magnetic field and the nucleus being observed. In general, there will be a very slightly greater proportion of spins aligned with the magnetic field than against it, depending on the sign of γ. This tiny difference in population means that NMR is intrinsically an insensitive technique. However, even this small population difference causes a bulk magnetisation aligned with the magnetic field, the magnitude of which is given by:

$$M_0 = \frac{N(\gamma h)2B_0}{16\pi^2 k_B T} \tag{2.6}$$

Applying Radio-Frequency (RF) pulses at the NMR frequency to the system perturbs the net magnetisation, rotating it by 90 degrees to make the net magnetisation perpendicular to the z axis. Once the RF pulse is turned off, the magnetisation precesses about B_0 and also decays back to equilibrium through various relaxation processes, such as spin-spin relaxation or spin-lattice relaxation, which will be explained further on. The precession is recorded by the spectrometer, and the resulting signal is known as the Free Induction Decay (FID). Performing a Fourier transform on the FID converts it into the NMR spectrum.

When placed inside a magnetic field, the electrons present around a particular nucleus induce a secondary magnetic field in opposition to the applied magnetic field. The strength of this induced field is dependent on the nuclear environment (the atoms directly connected to, or close to, the atom of interest), and causes chemical shielding. Each unique nuclear environment experiences a slightly different electron environment, which changes the energies of the quantised m_I states, shifting the Larmor frequency of that nucleus slightly. Chemical shifts, therefore, provide access to structural information. The chemical shift is a tensor property, and is usually anisotropic. In solution

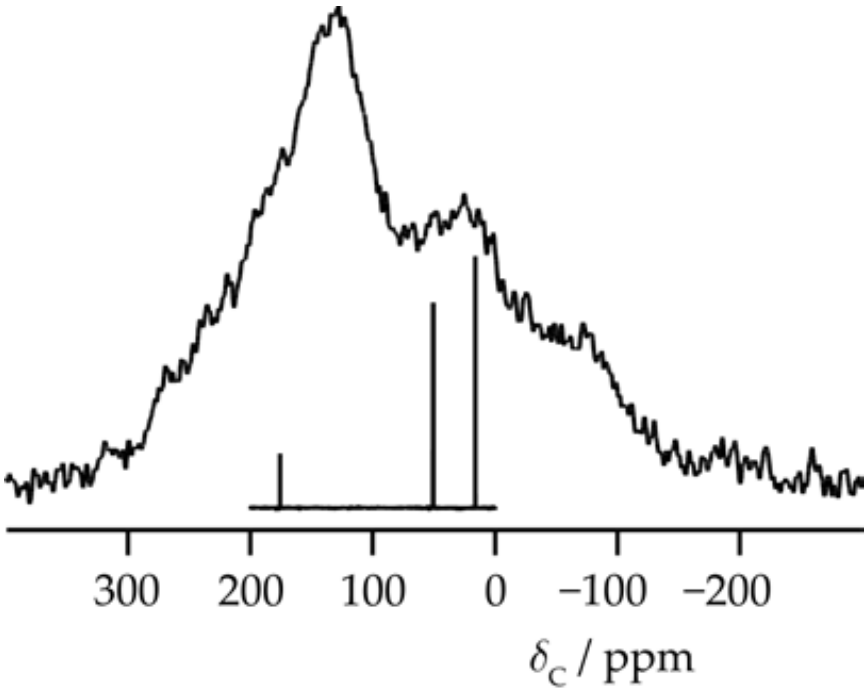

FIGURE 2.1: A typical Carbon-13 NMR spectrum of alanine. The top spectrum was recorded in the solid state, and the bottom in solution state.

state this causes little problem, as the motion of the molecules in solution mean this is often averaged out. In solid state, there is little motion, so the anisotropic properties must be taken into account. The shielding tensor links the induced magnetic field to the external magnetic field by:

$$\boldsymbol{B}_s = -\sigma \boldsymbol{B}_0 \tag{2.7}$$

The extent to which a particular atom is shielded depends on the angle between the external field and the molecular axis, θ, the angle having a specific effect of:

$$\frac{1}{2}(3\cos^2(\theta) - 1) \tag{2.8}$$

on the shielding. The anisotropic component of the chemical shift is called the chemical shift anisotropy (CSA).

The two magnetic field vectors have three components each (x, y and z), and

so the shielding tensor has 9 separate components: *xx*, *xy*, *xz*, *yy* ... However, the asymmetric contributions have very little influence on NMR spectra and can be safely ignored. Additionally, by choosing particular axes (known as the Principal Axis System or PAS), the symmetric components of the tensors can be made fully diagonal, and contain just three principal components: XX, YY and ZZ. In solutions, due to the rapid tumbling of molecules, these average to one isotropic component.

There are many interactions that are present in NMR experiments that can change the appearance of the NMR spectrum. Indirect (or J) coupling is associated with electrons co-ordinating their spins together. This occurs through bonds and allows molecular connectivity information to be obtained. J coupling is a prominent interaction in solution state NMR, leading to splittings in spectral peaks that is associated with covalently bonded neighbouring atoms. Analysis of the splitting patterns allows structural information to be gained. However, it is relatively unimportant in solid state NMR, as it is usually the smallest interaction with average values ranging from 10s Hz for light elements to 10s kHz for heavy metals.

The second interaction is dipolar coupling. This can be described as the interaction of one nucleus' dipole with another nucleus' dipole. For two different nuclei i and j, the energy relating to this interaction is expressed as:

$$E = \frac{1}{2}(3\cos^2\theta - 1)d_{ij}m_i m_j \tag{2.9}$$

where θ is the angle between the vector linking the two nuclei and the external field, and d_{ij} is the dipolar coupling constant, expressed as

$$d_{ij} = -\hbar(\frac{\mu_0}{4\pi})\frac{\gamma_i\gamma_j}{r^3} \tag{2.10}$$

which corresponds to the value of D_{zz} in the PAS. This means that the isotropic contribution of the dipolar coupling averages to zero, so dipolar coupling interactions cannot be seen in solution-state spectra.

Dipolar coupling in solids broadens the spectral line significantly, which can cause issues in identifying particular resonances. To alleviate this, decoupling techniques are often applied. The simplest decoupling technique is continuous wave decoupling, which involves applying a high power RF pulse at the ^{1}H frequency continuously while simultaneously acquiring the ^{13}C NMR signal. However, CW decoupling is relatively inefficient at moderate to high spinning speeds, and can damage the probe if applied incorrectly. Several other decoupling techniques use less power, but improve the efficiency in various ways, with the review by Hodgkinson giving an overview of many of them [84].

If the value for I is greater than $1/2$, quadrupolar coupling is present, which is affected by the distribution of charge around the nucleus. This nuclear electric quadrupole moment interacts with the local electric field gradient (EFG), which affects the energy levels of the nuclear spin and the NMR transition frequencies. The quadrupole coupling tensor shares similarites with dipolar coupling: it is traceless, and thus is isotropic and has no effect on solution state spectra. The nuclear EFG is a result of the local electron density, which depends on the symmetry of the nucleus.

The quadrupolar coupling affects each of the nuclear energy levels by:

$$E = \frac{3m_I^2 - I(I+1)}{8I(2I-1)}[(3\cos^2\theta - 1) + \eta\cos^2\phi\sin^2\theta]\chi \tag{2.11}$$

where θ and ϕ are spherical polar angles that describe the orientation of the local EFG PAS to the external field, and χ and η are the *nuclear quadrupolar coupling constant* and the *quadrupolar asymmetry* respectively.

Unlike in spin-half systems, in quadrupolar systems I is greater than $\frac{1}{2}$, and as such there are multiple transitions available between energy levels. This combines with the effect quadrupolar coupling has on those same energy levels, and shows up as multiple peaks in the NMR spectrum. However, if I is an **odd** multiple of $\frac{1}{2}$, the central transitions are largely unaffected due to the m_I^2 dependence in 2.11. This results in the central transitions remaining unshifted and narrow while the rest of the transitions in the powder pattern are broadened considerably.

2.3 MD Details/Methodology

Molecular dynamics is a method by which the motions of atoms and molecules within a given system can be simulated and 'observed'. There are a variety of methods within the remit of Molecular Dynamics, but most have the same principle: solving Newton's equations of motion to calculate the positions of atoms after each timestep. In this fashion, MD simulations are deterministic: given two points in a trajectory and the equations relating to the atoms, you can work from either point to the other and reach the same conclusion. This allows the time-dependent properties of systems to be studied, as well as statistical properties, making MD a powerful tool in the computational chemist's arsenal.

Molecular Dynamics is an umbrella term which covers a vast array of methods, each of which uses variations of the overall theme to observe different systems sizes and trajectory lengths. Performing *ab-initio* calculations allows the greatest level of accuracy to be obtained, but is only viable for small system sizes (around 100s of atoms) and short trajectories, with multiple picosecond trajectories being common for this level. Predicting transition states or excited states is the most common use for this calculation [85–87].

The next highest level of detail is atomistic models. These use force fields to represent the atoms as balls and the bonds as springs; and has a list of parameters and equations that describe how the atoms in the system interact with each other. In this fashion, the force on each atom is calculated, which is then used to calculate the position of the atoms after a given time, and the process is repeated over and over until a trajectory of suitable length is obtained. This method allows 1000s of atoms to be simulated, and trajectories of hundreds of nanoseconds are common, with microsecond trajectories also possible at the upper end.

Simplifying the atomistic simulations are the coarse-grained methods. In these simulations, groups of atoms are represented by a single ball, with simplified equations relating their overall interaction to other grouped atoms. These allow systems of tens to hundreds of thousands of atoms to be simulated, and very long simulations to be performed. However, the level of accuracy obtained from these simulations can be low unless extensive parameterisation is undertaken.

This work is primarily concerned with atomistic simulations, as these strike a balance between accuracy, detail, simulation length and ease of use [88–90]. While solid state simulations have been successfully completed using *ab-initio* techniques, these simulations are often on the order of ps [91, 92], whereas the motions we are interested in are on the order of ns or longer.

Coarse-graining the systems would allow us to extend the simulation time scale significantly, with large-scale biological systems having been simulated on the order of ms [93]. However, coarse-graining discards much of the individual detail of the molecules, as representing the system as peudo-atoms restricts the ability to extract these details. Additionally, coarse-grained force fields are optimised and dedicated to the simulations of biological systems [94, 95], making them unsuitable for our systems of interest.

2.3.1 Atomistic Simulation Details

At their heart, atomistic simulations involve solving Newton's equations of motion to predict the positions of atoms given the force acting on them and a fixed change in time. As stated before, the atoms (both nuclei and electrons combined) are represented as balls, while the bonds between are represented as springs. Each atom has a constant radius and constant charge throughout the simulation. Each spring will have a particular stiffness which represents the capacity of the bond to stretch or move, and a particular length which represents the equilibrium distance between the two atoms. This list of parameters: bond length, bond strength, charges and charge interactions are combined into a force field.

2.3.2 Force Fields

The interactions between the atoms in an atomistic simulation are collected into a force field. These interactions are added together to obtain the potential energy of each atom and the system as a whole. The general form of the force field is:

$$E_{\text{tot}} = E_{\text{bonded}} + E_{\text{non-bonded}} \tag{2.12}$$

While each term in that general form can be broken down further into:

$$E_{\text{bonded}} = \Sigma E_{\text{bonds}} + \Sigma E_{\text{angles}} + \Sigma E_{\text{torsion}} \tag{2.13}$$

$$E_{\text{non-bonded}} = \Sigma E_{\text{VdW}} + \Sigma E_{\text{electrostatic}} \tag{2.14}$$

Where E_{bonds} relates to the bond length, E_{angles} to the bond angles, E_{torsion} to the dihedral angles etc. Once the energy of the interactions is calculated by referencing the current positions of the atoms and the force field parameters, the derivative of this is taken with respect to the current atomic positions. This obtains the force acting on each atom, which can be used to calculate

its new position and momentum, and the simulation moves forward one timestep. Most simple bonded force field equations are of the form:

$$E_{\text{bonded}} = \Sigma k_i (x - x_0)^2 \tag{2.15}$$

Where E is the energy, k_i is the force constant of this 'spring' and x and x_0 are the current and equilibrium position respectively. Bond lengths and bond angles are most commonly treated in this way, with the interactions being treated as a simple spring under Hooke's law: any deviation from the equilibrium position produces a force which attempts to eliminate that deviation. The exception to this are torsion angles, which use a different equation due to their periodic nature:

$$E_{\text{torsion}} = k_{\text{torsion}}[1 + \cos(n\theta - \theta_0)] \tag{2.16}$$

Where θ is the torsion angle, n is the multiplicity of the rotation (how many identical orientations there are), and θ_0 is the angle at the height of the energy barrier. Additional terms can be introduced into these equations to increase the accuracy of the simulations, albeit at a higher computational cost per timestep. The non-bonded parameters (electrostatics and van der Waals forces) are based on the partial charges of the atoms in the system. Calculating these parameters is a key step in the simulation, as it affects both intermolecular and intramolecular interactions. The simplest model of the electrostatic interactions is given by Coulomb's law, which assumes that the electrostatic charges are point charges centered on the nuclei, and is expressed by:

$$V_{\text{elec}} = \sum_{n_i} \sum_{n_j} \left(\frac{q_i q_j}{\epsilon r_{ij}}\right) \tag{2.17}$$

Where $q_{i/j}$ represents the charges on atoms i and j, r_{ij} is the distance between the atoms and ϵ is the dielectric constant of the medium between the two.

The generation of force field parameters requires determining the equilibrium points for all bonded interactions and determining the partial charges and polarizability of atoms for the non-bonded interactions. This process can involve successive geometry optimisations using *ab-initio* methods at increasingly accurate levels of theory, before using additional pieces of software to calculate and extract the required properties. This iterative method is used as performing a geometry optimisation at a high level of accuracy from a completely unoptimised starting configuration can be prohibitively long. By using a lower accuracy but faster method to begin with, we approach a minimised structure rapidly, and can then use increased levels of accuracy to fine tune our geometry. As this can be costly both in computational terms and in real-world time terms, most force fields for particular systems are adapted from commonly used force fields, with only the additional parameters required for this system added to the overall force field. In this work, the CHARMM36 force field [96] was used for the simulations.

The CHARMM force field was initially designed for use with the CHARMM molecular dynamics program, but the force field has been released for general use as well. In particular, a general topology generator for drug-like molecules was released: CGenFF [97, 98]. A topology generator produces a topology for a molecule, that is, an assignment of each atom in a molecule to an atom type in a particular force field, detailing the interaction parameters to be used during a simulation. This covers a wide range of potential systems, and force fields can be easily generated using the online tool. Additionally, the force fields come with a 'penalty score', detailing the accuracy of the parameters given for each interaction. This is generated by comparing the geometry of the submitted molecule with the CGenFF reference geometries, and picking the most appropriate interaction. The extent of deviation from the most appropriate interaction is represented by the penalty score.

CGenFF is suitable for small molecule simulations, where the interactions are generally simple, or for protein simulations, which have been heavily parameterised. For molecules such as furosemide, the generator can only produce a best guess, leading to high penalty scores. For this reason, generation of parameters for complex, uncommon molecules is best done via an alternative method.

2.3.3 Integration Algorithms

In a simulation, the equations of motion are solved by a procedure known as the finite difference algorithm, where the positions and dynamic information of the system at time t is used to calculate the positions and dynamic information of the system at time $t + \delta t$. The equations of motion for a simple system are given by:

$$M_i \frac{\delta^2 r_i}{\delta t^2} = F_i \tag{2.18}$$

$$F_i = -\frac{\delta E}{\delta r_i} \tag{2.19}$$

Where F_i is the force on the atom, m is the mass, t is the time, E is the potential energy (generated from the force field parameters) and r the atomic coordinates. We can then use the force field to calculate the gradient of the potential energy over position, then use that as the force to calculate the acceleration of the atoms. The coupled equations above cannot be solved analytically, but can be calculated numerically, to achieve a set of new positions for each atom. There are several different algorithms developed for this purpose.

Verlet Algorithm

The Verlet algorithm was developed by Verlet in 1967 [99] and is the starting point for most modern integration methods. A third order Taylor expansion is used to calculate the positions of the atoms both forwards and backwards

in time:

$$r(t+\delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 \tag{2.20}$$

$$r(t-\delta t) = r(t) - v(t)\delta t + \frac{1}{2}a(t)\delta t^2 \tag{2.21}$$

Where r is the position, v is the velocity and a is the acceleration. By combining the two expressions, an expression for the positions at t+δ t is found:

$$r(t+\delta t) = 2r(t) - r(t-\delta t) + a(t)\delta t^2 \tag{2.22}$$

This algorithm is computationally cheap, requiring low data storage, however velocities are not generated. Improvements to this algorithm have been proposed and widely implemented.

Leap-Frog Algorithm

The leap-frog algorithm calculates the velocities at a half time-step after the reference time t using the equation:

$$v(t+\delta t/2) = v(t-\delta t/2) + a(t)\delta t \tag{2.23}$$

These velocities are then used to generate the atomic positions at time t+δt:

$$r(t+\delta t) = r(t) + v(t+\delta\frac{t}{2})\delta t \tag{2.24}$$

This algorithm is called the leap-frog algorithm as the velocities 'leap' over the positions which then leap over the velocities etc.. Velocities are calculated accurately as a result of this method, but only in terms of the half time-step, meaning they are not calculated at the same trajectory time as the positions. If this is required, the velocity is averaged from the $+\frac{1}{2}$ and $-\frac{1}{2}$ times. This algorithm can be easily applied to complex molecules and is often the first choice for molecular simulations.

Velocity Verlet Algorithm

The Velocity Verlet algorithm works in much the same way as the regular Verlet algorithm, albeit with an extra step. Once the particular positions and forces are propagated from the current time step, the algorithm uses a combination of the current and future accelerations to calculate the velocity at the timestep. While this increases the precision of the simulation, it does so at a higher computational cost.

Choosing which algorithm to use is a key step in the production of a suitable MD trajectory. The software used, GROMACS, utilises the leap-frog algorithm preferentially over the velocity verlet, and so this was chosen. As the simulations will be grouped into states, achieving a very high level of accuracy is not necessary, and the size of the systems to be simulated means that a significantly higher computational cost would need to be paid.

2.3.4 Thermodynamic ensembles

Molecular dynamics is performed on microscopic simulations in an attempt to mimic the behaviour of macroscopic systems. In order to achieve this, we have to use a thermodynamic ensemble, a collection of systems that have different microscopic states but the same overall macroscopic or thermodynamic state. Various ensembles are used in simulations:

- **Microcanonical:** The thermodynamic state has a fixed number of atoms, N, a fixed volume, V and a fixed energy, E, and is an isolated system.
- **Canonical:** This state has a fixed number of atoms, a fixed volume, and a fixed temperature, T.
- **Isothermal-isobaric:** This state has a fixed number of atoms, a fixed pressure, P, and a fixed temperature.

- **Isoenthalpic-isobaric:** This state has a fixed number of atoms, a fixed pressure, and a fixed enthalpy, H.
- **Grand canonical:** This state has a fixed volume, a fixed temperature, and a fixed chemical potential, μ.

MD simulations can be performed in any of these ensembles, with the default being microcanonical. However, it is often in our interest to use the canonical or isothermal-isobaric ensembles to extract different data.

Temperature coupling

Temperature coupling algorithms are used to control the temperature of the system. As the temperature is a function of the average kinetic energy of a system, we can control the temperature by scaling the velocities of all atoms accordingly. Various methods have been developed for this, one of which is the Berendsen thermostat [100]. This algorithm mimics the weak coupling of the simulation system to an external heat bath at the desired temperature, T_0. The deviation of the system temperature from T_0 is given by:

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau} \tag{2.25}$$

where τ is a time constant. This method suppresses fluctuations in the kinetic energy of the system, and so does not generate a true canonical ensemble. This affects the results for very small systems, or when calculating the kinetic energy of the system. The velocities become rescaled by a scaling factor, λ, given by:

$$\lambda = [1 + \frac{\Delta t}{\tau_t}(\frac{T_0}{T(t - \frac{\Delta t}{2})} - 1)]^{\frac{1}{2}} \tag{2.26}$$

where τ_T is close, but not equal to the time constant, τ. In practice, this scaling factor is limited to avoid very large scaling at any individual time step, which could cause a failure in the simulation.

The velocity rescaling thermostat [101] is similar to the Berendsen thermostat, but allows production of the correct ensemble by adding a random force to the system.

These two algorithms are weak coupling algorithms, used to relax the system to a desired temperature. In order to probe the correct canonical ensemble, the Nosé-Hoover temperature coupling [102] is frequently used. The main features of this algorithm are the introduction of a thermal reservoir and a friction coefficient, allowing for kinetic energy changes in the system to be observed.

Pressure coupling

In a similar vein to temperature coupling, we can also couple the system to a pressure bath. Using the Berendsen algorithm is the simplest solution, which rescales the coordinates and box vectors of the system at every step towards a set reference pressure, P_0, so that

$$\frac{dP}{dt} = \frac{P_0 - P}{\tau_p} \tag{2.27}$$

Rescaling is performed by appling a scaling matrix, given by:

$$\mu = \delta_{ij} - \frac{\Delta t}{3\tau_P}\beta_{ij}[P_{0ij} - P_{ij}(t)] \tag{2.28}$$

where β is the isothermal compressibility of the system.

In certain systems, fluctuations in the pressure are important, and as the Berendsen algorithm is only weak coupling, this cannot generate the exact

thermodynamic ensemble. By using the Parrinello-Rahman pressure coupling [103], which is similar to the Nosé-Hoover coupling for temperature, we can generate the correct thermodynamic ensemble. The Parrinello-Rahman coupling works by updating the equations of motions for particles at every step.

2.3.5 Equilibration

Before a system can be simulated and useful data extracted, the initial system needs to be equilibrated. The equilibration allows the system to lower its energy, as often a starting set of coordinates is not in the optimal positions, so giving time for the system to equilibrate is crucial. This equilibration can also allow the system to spread out among the states available to it, something which is of key interest in this work.

Equilibration also needs to occur after a system has been heated to a desired temperature. This steps allows the kinetic and potential energies to equilibrate specifically, allowing the kinetic energy added during heating to spread out among the available degrees of freedom. Equilibrating the system is reasonably simple: The system is simply simulated for a period of time using a weak coupling algorithm, typically Berendsen, so the energies can exchange through the degrees of freedom. Plotting a graph of the potential energy and waiting for this to level out indicates that equilibration is complete.

2.3.6 General Simulation Details

A set of common simulation practices were used throughout the entirety of this work. While specifics for each system have been given in the following chapters, the basic outline common to each will now be detailed.

Initial starting positions were obtained from CIF files, either accessed via the CCDC [104] or using internally generated CIF files. These starting coordinates were typically comprised of a unit cell, and the cells were multiplied to create a super cell of varying sizes, as specified in the following chapters. Force field parameters were obtained using the CGenFF tool for the diamondoids, Chapter 3, or the AMBER force field in the case of the FSPA system, Chapter 4.

All simulations were run using the GROMACS 2016.4 molecular dynamics software [105–111]. Starting coordinates were first energy minimised to within 100 kJ mol^{-1} nm^{-1}, and then equilibrated for 1 ns. Systems then underwent an annealing run to achieve the correct simulation temperatures, usually increasing the temperature every 20 ns of simulation time. These preparatory simulations were run using the Berendsen thermo- and barostat. Deviations from the usual temperature increase have been specified in the following chapters.

Once snapshots of particular temperatures had been obtained, they were equilibrated for 10 ns before undergoing the production run. The productions runs switched to using the Parrinello-Rahman thermostat set at the appropriate temperatures, and the Nose-Hoover barostat. Anisotropic pressure couplings were used, to allow each system some flexibility in their motion. Some simulation boxes twisted out of shape because of this, and so the off-diagonal components of the pressure coupling were set to 0, to ensure a rectangular box in these cases. This has been discussed in greater detail in section 3.3.2.

2.4 Principal Component Analysis

Principal Component Analysis (PCA) is a procedure that converts a set of observables of potentially correlated variables into a set of linearly uncorrelated variables known as Principal Components. The first principal component generated has the greatest variance, with each subsequent component having the next greatest variance while also being orthogonal to the preceding components.

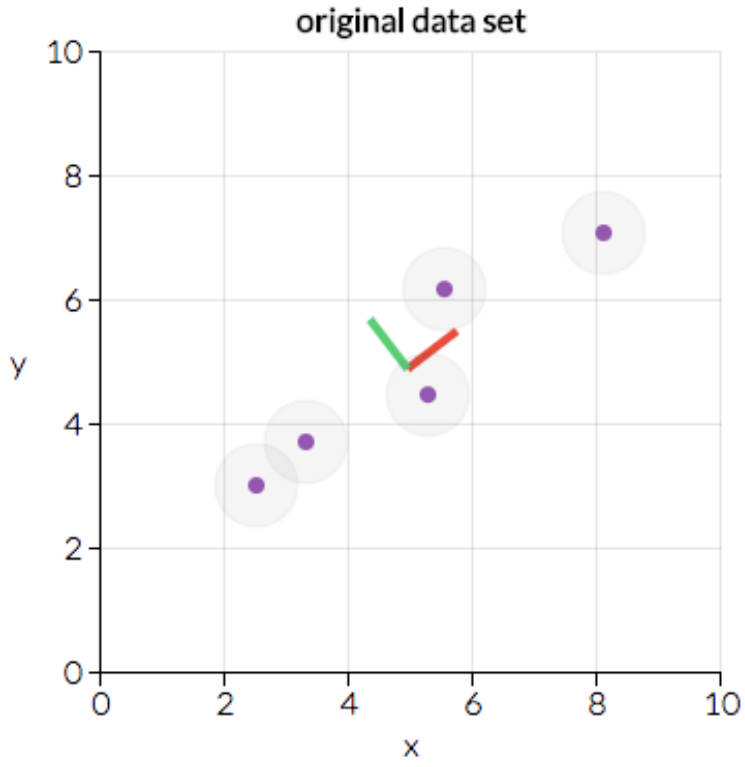

FIGURE 2.2: Example data points showing the first (red) and second (green) principal components.

Figure 2.2 shows an array of data points along a coordinate axis. To find the first principal component, a straight line which has the greatest spread of data when the points are projected onto it needs to be found. In this case, that straight line is shown in red in figure 2.2. This line is defined as the first principal component or the first eigenvector. A second principal component

can be found, under the constraint that it is orthogonal to the first, and is shown in green.

The result of this is a series of eigenvectors and eigenvalues. The eigenvectors are the direction of the lines drawn, showing which direction we have to 'move' in to achieve the greatest variance. The eigenvalues associated with each eigenvector is simply the amount of variance associated. The eigenvector with the highest eigenvalue is therefore the first principal component, with each eigenvector ranked in terms of its eigenvalue.

The number of possible eigenvectors is equal to number of dimensions the data set has. PCA does not alter the data, it simply transforms it into a new set of coordinates which showcase the maximum variance along each axis. In our previous data, we have two dimensions: X and Y so only two possible eigenvectors. In a 3D data set, we have three dimensions: X, Y and Z, so three possible eigenvectors, and as we increase the number of available dimensions, the number of eigenvectors increases accordingly. Figure 2.3 shows this transformation.

2.4.1 Time-lagged Independent Component Analysis

Time-lagged Independent Component Analysis (TICA) is a similar method to PCA, taking in a large data set and finding components which can describe the data in a different way. The key difference is that TICA uses eigenvectors which maximise the autocorrelation of the data rather than the variance. When we maximise the variance, we look for the biggest changes in state, as these will give the greatest difference in values. When we maximise the autocorrelation, we look for the slowest changes in state, as a high autocorrelation means the system remains in a similar state for long periods of time. TICA also lifts the constraint that each component must be orthogonal to the

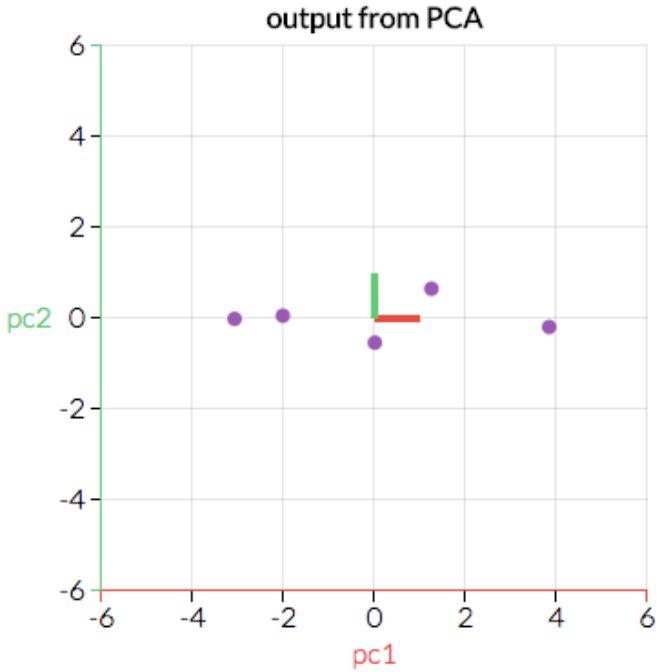

FIGURE 2.3: The reoriented data with respect to the principal components.

preceding components. Figure 2.4 shows a comparison between PCA and TICA.

2.5 MSM Theory and Application

At its most basic, a Markov State Model (MSM) is used to model randomly changing systems, where all possible future states are assumed to depend only on the current state, and not on the history of the system. For systems that obey the Markov Property i.e. are memoryless, a sufficiently detailed MSM can be used to detail the system exactly. A MSM can also be called a Markov Chain where the system is fully observable and autonomous. Markov chains were first studied by Andrey Markov, with his first paper on them published in 1906 [112].

A MSM can be best represented by a transition matrix. One edge of the matrix represents the state of the system at time t, and another edge the state of

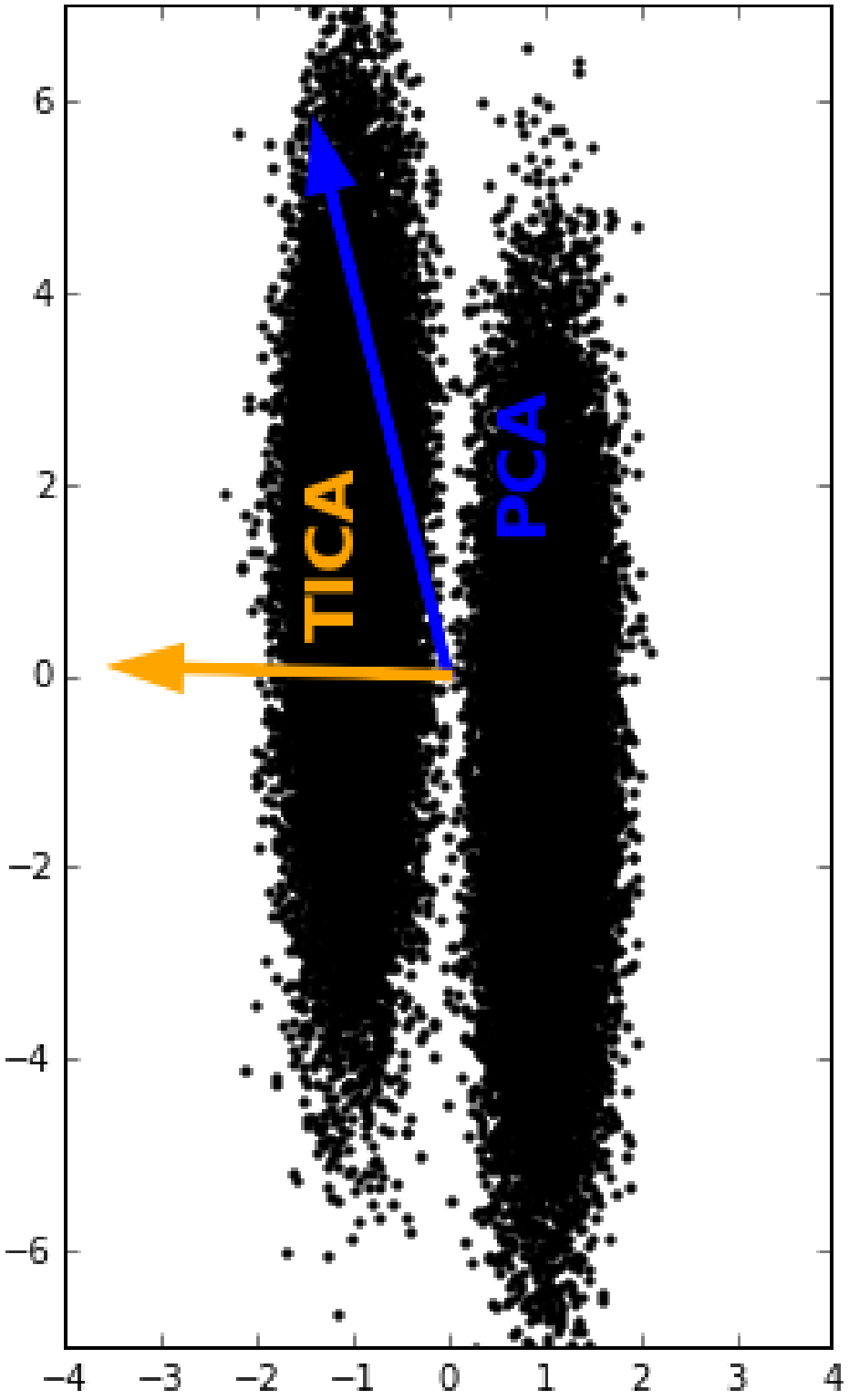

FIGURE 2.4: Comparing PCA and TICA for the same dataset. The arrows indicate the first component when clustering the data using each technique.

the system at time $t+\delta t$. The values of the matrix represent the probability of moving from any state a at time t to any other state (including the starting state) b at time $t+\delta t$. 2.5 shows an example of such a matrix.

$$P = \begin{bmatrix} P_{1,1} & P_{1,2} & \dots & P_{1,j} & \dots & P_{1,S} \\ P_{2,1} & P_{2,2} & \dots & P_{2,j} & \dots & P_{2,S} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ P_{i,1} & P_{i,2} & \dots & P_{i,j} & \dots & P_{i,S} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ P_{S,1} & P_{S,2} & \dots & P_{S,j} & \dots & P_{S,S} \end{bmatrix}.$$

FIGURE 2.5: General Markov matrix.

A key idea of MSMs is their time-independence. The Markov property already states that future states only depend on the current state, not on the states that have occurred before. Any data or analysis derived from the model is therefore independent of the current time of simulation, as the probabilities of moving from the current state to any possible state is fixed.

Constructing a MSM requires several pieces of data: first, a list of the possible states the system can adopt. The MD simulations performed give a trajectory of the positions, velocities and accelerations of the atoms of the system at a given time, but feeding each of these in as a different state would produce a MSM that simply recreates the simulation, and doesn't allow us to predict new events, as well as being prohibitively large. To alleviate this, we can featurise the trajectories: extract particular pieces of information about the molecules to be used to feed into the MSM generator.

These features can vary: in protein simulations, the backbone is of particular interest, and so the torsion angles, positions, or key distances of these atoms

can be used as the features. For the systems in this thesis, intramolecular vectors were used as the features, as these give us key information about the orientation of the molecule. Specifying a pair of vectors perpendicular to each other allows a complete description of the orientation of the molecule to be found.

However, simply entering the full set of possible vector values into an MSM generator would still only recreate the original trajectory. To solve this, we need to discretise the featurised trajectory into groups of common states. To do this, we introduce the use of clustering algorithms.

2.5.1 K-Means Clustering

The K-means clustering algorithm [113] has been used extensively during the data analysis. Clustering algorithms in general all have the same objective: to group objects together in a way that means objects within one group are more similar to each other than to those in other groups. An example would be clustering numbers into discrete integers: a set of numbers $1.8, 2.1, 3.7, 4.2$ could be clustered into two groups: 2 and 4. The main purpose of this to to allow meaningful comparisons to be drawn: Comparing each data point in a data set with every other point would take an extraordinary amount of time, but making comparisons between the populations of clustered groups allows relationships and correlations to be drawn.

K-means clustering is a clustering algorithm which aims to partition a dataset in k clusters, with each data point belonging to the cluster with the nearest mean, which then serves as the cluster centre or value of that cluster. As a result of this, the data space is partitioned into Voronoi cells, areas of space differentiated by their closest cluster centre. Formally, the K-means clustering algorithm (also known as Lloyd's algorithm [114]), aims to cluster a given

set of observables into n sets to minimise the within-cluster sum of squares or variance within a cluster. This can be expressed as:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}\in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i \tag{2.29}$$

where $\mathbf{x}$ is an observation, S_i is a set grouping observations together, $\operatorname{Var} S_i$ if the variance within that set, and $\bar{\mathbf{i}}$ is the mean of all the points included in S_i.

2.5.2 Markov Model creation

Once the cluster centres, μ_i, have been created, the trajectory is then discretised. This simply involves finding which cluster centre the system is closest to at each point in time, and replacing the trajectory data with the number of the state. This is then the basis for the construction of the Markov model. To construct a Markov model, we count the number of transitions from every state a to any possible state b, including the starting state. These transition numbers are then converted to the probability of going from state a to state b by dividing by the total number of transitions. The result of this is a transition matrix, as discussed above.

Counting transitions seems like a relatively straightforward process, however in practice the amount of data we have is a limiting factor. In an ideal example, where we have a large excess of data, we can look at transitions 'end-to-end': specifying a lag time τ, we note the transition between states at time 0 and time τ, then between time τ and time 2τ and so on. This is the independent counts method of transition counting. However, with a finite amount of data, this leads to imprecise estimates of the transition probabilities, as well as discarding large amounts of data.

Instead, we can use a sliding window method of counting. In this method, we start at time 0, with the same lag time τ, but instead of moving our start point in multiple of τ too, we simply move to the next point in the trajectory. For example, if $\tau = 5$, we would note the transition between time 0 and 5, 1 and 6, 2 and 7 etc. In this way, we increase the precision of the transitions probabilities, but at the cost of underestimating the model uncertainty.

2.5.3 Perron Cluster Cluster Analysis

Perron Cluster Cluster Analysis (PCCA), is a method for constructing coarse-grained Markov models by using the eigenspectrum of a transition probability matrix. The term "Perron Cluster" relates to a set of eigenvalues that are close to the largest eigenvalue within a particular range, and have a reasonable gap in values between them and the rest of the set of eigenvalues. For example, if an eigenspectrum consisted of: 1, 0.99, 0.98, 0.5, 0.45, 0.1, then three Perron Clusters could be constructed: 1, 0.99 and 0.98 as one, 0.5 and 0.45 as another, and 0.1 as the last. Once we have obtained these Perron Clusters, we can then use the eigenvectors to coarse-grain our MSM.

We can convert the eigenvalues of a transition matrix to a series of time scales. The corresponding eigenvectors describe which transitions are occurring on the timescales. The first and largest eigenvalue λ_1, is always 1 for a model that is connected and at equilibrium. The components of the associated eigenvector are the equilibrium populations of each state. The remaining eigenvalues λ_n, where $n > 1$, have right eigenvectors (column vectors) that describe the interconversion between states, whereas the left eigenvector (row vector) contains the same information but weighted by the equilibrium population.

The PCCA method begins by converging all microstates (the clusters obtained from the k-means clustering) into one macrostate. It then consults the

slowest right eigenvector and splits this initial macrostate into two, based on the components of the microstates in the eigenvector. For example, consider a 4 centre clustering that we wish to split into 2 macrostates. Consulting the slowest right eigenvector tells us that states 1 and 3 have a less than or equal to 0 component in this eigenvector, whereas states 2 and 4 have a positive component. This means we split the centres into 2 macrostates: 1 and 3 as one macrostate, and 2 and 4 as the other.

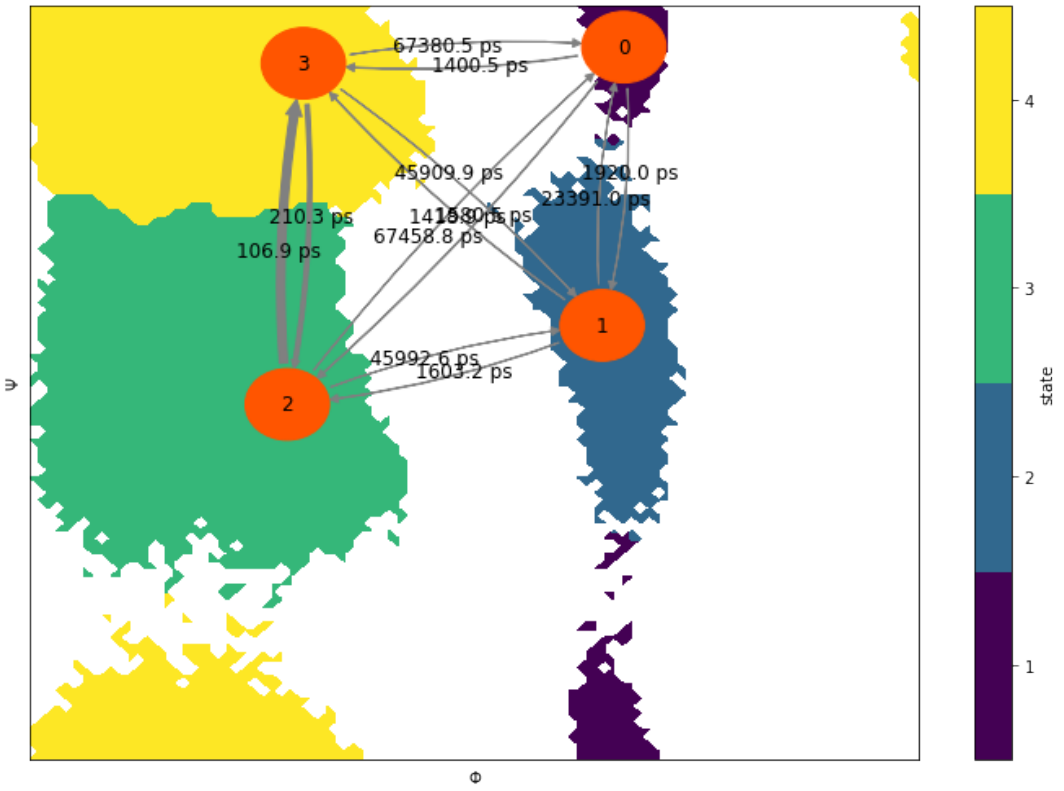

FIGURE 2.6: Example of the PCCA method. States 1 and 3 have a 0 or lower component in the slowest right eigenvector, whereas 2 and 4 have a positive component. As such, we would split the states into two macrostates: 1 and 3, and 2 and 4.

One useful feature of this method is that it provides a simple way to determine the number of macrostates. If the system has a well defined set of free-energy wells, then the system should show a clear separation of timescales. In general, if there is a gap after eigenvalue n, then you can construct a reasonable macrostate model using n macrostates.

2.5.4 Hidden Markov Models

This section will introduce Hidden Markov models. This is a vital part of the analysis, as producing a HMM from the data is the method that allows us to group similar cluster centres together, and then extract these overall states and transition times from the original data.

A Hidden Markov Model (HMM) is a Markov Model which contains states that are unobserved or 'hidden' [115]. In a regular Markov Model, we can observe all the states the system can adopt, and transitions between these states are direct. In a HMM, the underlying states are not directly observable, but output of the states is observable. Each underlying state has a probability distribution of causing a particular observable output to be seen, and so by counting transitions from these different outputs, we can infer some information about the sequencing of the underlying states. This is illustrated in 2.7

With respect to this work, construction of a HMM is based upon the PCCA analysis described in section 2.5.3. The macrostates are taken as the observables outputs, with the microstates taken as the underlying states. This permits us to perform several actions: Firstly, we can calculate the time of going from observable to observable by grouping the microstates together. Calculating this time by using small numbers of clusters both increases the approximation error and can lead to incorrect calculations of transitions if a particular molecule "hovers" about the border between two clusters. Secondly, we can obtain a set of properties for the observable which corrrespond to the average orientation of the molecule. This allows us to "see" which orientations the molecule moves between, which gives us insight into the motion.

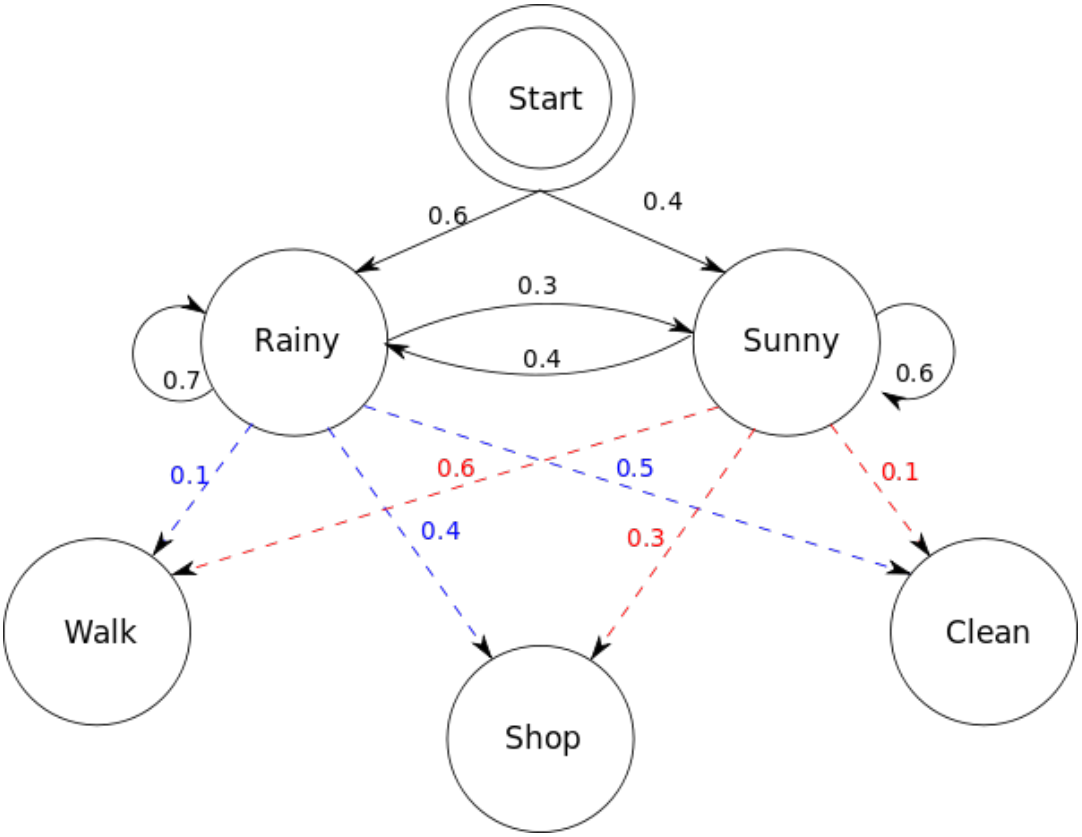

FIGURE 2.7: Example of a hidden Markov model. In this, we hear about what a person did during their day, and can make estimations of the weather based on that. The state the weather is in is the unobserved state, whereas the activity a person undertook is the observable. By using our knowledge of the transitions between the unobserved states and the likely activity based on that, we can estimate the probabiliy of the weather being either rainy or sunny based on the activity performed.

2.6 Use of Python and the PyEMMA module

The PyEMMA module [116] is a Python library designed for the estimation, validation and analysis of MSMs based on data obtained from MD simulations. This library has been the main method by which MSMs on the data have been developed, with various supplementary code utilising the methods tailoring the input and output data to our needs. As well as containing the code required to construct MSMs, it also contains a variety of data clustering and discretization tools, which are also essential to creating MSMs. In particular, the code has the ability to perform PCA and TICA analysis on an

arbitrary data set provided to it.

The PyEMMA module was originally created with proteins in mind, but many of its tools can be adapted for general use. As discussed in section 2.5, rather than loading in trajectory files and using the module's selection tools, we perform the vector extraction prior to loading the data in. This ensures that we are analysing only the relevant data of interest, and increases the speed of analysis by reducing file sizes. Once this data has been loaded in, the next step is to capture the important dynamics by performing a PCA or TICA analysis upon it. As stated before, performing a TICA analysis requires the input of a suitable lag time. The choice of lag time comes from calculating implied timescales at several different lag times, and choosing the time that allows the timescales to converge at the earliest opportunity.

2.6.1 Implied Timescales

During the course of the analysis, we create implied timescale plots: these are plots of the relaxation timescales of a molecule implied by a Markov model that has been estimated at a particular lag time. In creating these plots, we generate several models at a variety of lag times, usually seperated by a factor of 1.5, and show how the relaxation timescales vary as you vary the lag time. Each of these plots contains a greyed out area: this area has a timescale that is shorter than the lag time chosen to generate the Markov model. If a timescale is within this area, it indicates that the time between measurements for the Markov model is greater than the timescale of the relevant process, and so details of the process will be lost.

The implied timescale plot is a general test of the Markovianity of the data, but is only an indicator of a good lag time to choose, as MSMs tend to underestimate the implied timescales. The timescales will steadily approach the

"true" timescale as the lag time increases, however increasing the lag time discards a greater amount of data. As the relaxation times for a Markov model with a lag time of τ should be equal to a model with a lag time of $n\tau$, we choose the smallest lag time available at which the model gives Markovian behaviour i.e. the smallest time at which the timescale is constant. This property can be expressed as:

$$T(n\tau) = [T(\tau)]^n \tag{2.30}$$

where T is the transition matrix, n is some constant, and τ is the lag time. Using the example figure 2.8, we can see that the Markov time is approximately X frames, so we can choose this as the lag time.

Once this lag time has been chosen, the resulting transformed data needs to be clustered. PyEMMA uses the k-means clustering as its method of choice, which is suitable for our purposes. However, choosing the number of clusters is one of the drawbacks of the k-means method. To determine the number to be entered, the transformed data is clustered repeatedly to a varying number of cluster centres. We look for the same criteria as choosing the lag time: the earliest convergence of the implied timescales. This ensures that motions of interest are modelled accurately, and choosing times that have earlier convergence allows us to see more of the relevant motions when creating the models.

Once the clustering is complete, the trajectory is discretised. This is a simple process that involves assigning each point in time in a trajectory to a particular cluster centre, giving us a list of 'states' over time rather than vectors. This is also a crucial point, as assigning the trajectory to states allows us to create the Markov model. First, the implied timescales of the discretised trajectory are calculated, and a plot of these created, the interpretation of which is discussed in section 2.7. Once a suitable lag time for the model is selected, the Markov model is constructed by PyEMMA, giving us an MSM object on

which we can perform further analysis.

Due to the often high level of clustering performed in the previous steps, extracting meaningful transitions between overall states requires us to group cluster centres together. This is achieved by using the PCCA+ method as described in section 2.5.3. To determine the number of macrostates, we plot a graph of the ratio between the n th and n th+1 timescales. As the timescales are directly calculated from the eigenvalues, we can see the separation between the eigenvalues, and choose an appropriate number of macrostates, with a ratio of over 1.5 indicating a good point to partition.

Once the number of macrostates has been determined, we can coarse grain the MSM into a hidden Markov model, as described in section 2.5.4. Once this has been performed, we can use the new clustering to extract representatives of the macrostates, allowing us to 'see' what position the molecules are in for each state. This also allows us to obtain the transitions rates from state to state, as well as inferring the mechanism of movement by following the path from one macrostate to another.

2.6.2 Transition Path Theory

During our analysis, we want to extract the rates and pathways of moving from a state A to a state B. As we may have multiple states within our system, it is of interest to determine whether transitions are direct from A to B or via any intermediate state C. In order to achieve this, we can utilise Transition Path Theory (TPT) with our hidden Markov model. The basics of TPT involve first splitting the states available to us into A, our start point/s, B, our end point/s, and I, our intermediate states. Choosing multiple start points allows us to cluster similar states together, similar to constructing an HMM from an MSM. We then calculate the probability of moving towards state B for every intermediate state.

A main consequence of these ideas is the reactive flux. As we are only interested in trajectories that start from state A and move to state B without going back to A beforehand, we can calculate the flux associated with these reactive trajectories by multiplying the probability of coming from state A by the probability of moving on to state B. However, we also want the quickest pathway from state A to B, without considering any detours or recrossings. We can then calculate the total flux between these two sets. This total flux tells us the number of expected transitions from set A to set B per time unit τ (the time step of the transition matrix) that an infinitely long trajectory would produce. However, such a trajectory would cycle between A and B, and would include the time taken to go from A to B and from B to A. To correct for this, we have to only count the forward trajectory segments.

These ideas have been implemented into the PyEMMA package, and have been used to calculate the transition times between states in the various systems. The idea of transition path theory was combined with Markov models by *Weinan* who developed and detailed the methods outlined above [4]. This was further developed with regards to Markov jump processes, jumping from state to state rather than moving along a continuum of state space [117]. This idea was also independently introduced in terms of reversible rate matrices, Markov model transition matrices where the probability of moving from state i to state j is equal to the probability of moving from state j to state i, by *Berezhkovskii* [118]. The various algorithms used to apply this theory within the PyEMMA module were introduced by *Noé* in 2009 [119], and are the algorithms used to apply this theory within this work.

2.7 Interpretation of Results

This section will go through how we interpret the results from the analysis method. It will detail what the free energy diagrams show, how to read an

implied timescale plot, how to decide on the number of metastable states to choose, how to extract useful transition data from the Markov matrix, how path theory works within the framework of the Markov model, how we extracted correlation coefficients for adjacent molecules and how to interpret them.

During the course of the analysis, a variety of graphs are produced. Some of these graphs have been touched on in previous sections, however this section will explain how to interpret each of these graphs correctly, and detail what we can do with this information.

2.7.1 Free Energy Diagrams

The free energy diagrams produced by pyemma, as exampled by figure 2.9, are essentially a Boltzmann inversion of the populations of each component across the length of the trajectory. The data is first converted into a histogram, which describes the probability of the system adopting a particular orientation at any given time. This is then converted into a free energy using the equation:

$$F(x) = -k_{\mathrm{B}} T ln H(x) \tag{2.31}$$

Where $F(x)$ is the free energy at x, k_{B} is the Boltzmann constant, T is the temperature and $H(x)$ is the histogram of the data. This results in a diagram that shows areas of low energy/high probability, as well as areas of higher energy between them. The lowest energy areas on the graph would be the states the system is able to adopt, as it spends the most amount of time here, with connecting areas forming the transition pathways from state to state. In figure 2.9, we can see three distinct areas of low energy, each of these corresponding

to a state that diamantane can adopt, with connecting areas of higher energy that form the transition pathway between states.

2.7.2 Metastable partitioning

As discussed in section 2.5.3, the eigenvalues of the model can be used to determine how many metastable states we can divide the model into. Simply plotting a graph of the magnitude of each eigenvalue should give a general idea by eye of the number of states, however taking the ratio between an eigenvalue i and the following eigenvalue $i+1$ gives us a numerical way of determining this. An example plot is below, as figure 2.10.

2.7.3 States and state transtions

The previous steps have allowed us to reduce the data down to a more manageable set, identify common states among the trajectories, and then group these to determine the metastable states present in the system. However, we cannot go directly from the TIC values to the metastable states. To allow us to obtain an example of each state from the trajectory, we take several steps. First, we recluster the data, but specifying a number of cluster centres equal to the number of metastable states. We then find the data point that is closest to each of the cluster centres, and work backwards to find both the molecule and time step that point corresponds to. From this, we can view its orientation from the trajectory and "view" the state directly.

2.7.4 Correlation of Motion

The motions we have been considering so far have been isolated motions, treating each molecule as if disconnected from the others. However, each

molecule exists with others in a system, and so there may be correlation between the molecules' motions. Using the coordinates to determine this correlation would be extremely difficult, however we can use the dimension reduction methods described above. By reducing the data from raw coordinates to the TICs or PCs, we can then calculate the correlation between each molecule's component in these TICs over time, and obtain a correlation coefficient. High coefficients will suggest that the two molecules are highly correlated, and that when one moves or twists in on fashion, so does the other, whereas low coefficients will indicate that the two molecules are essentially random with respect to each other.

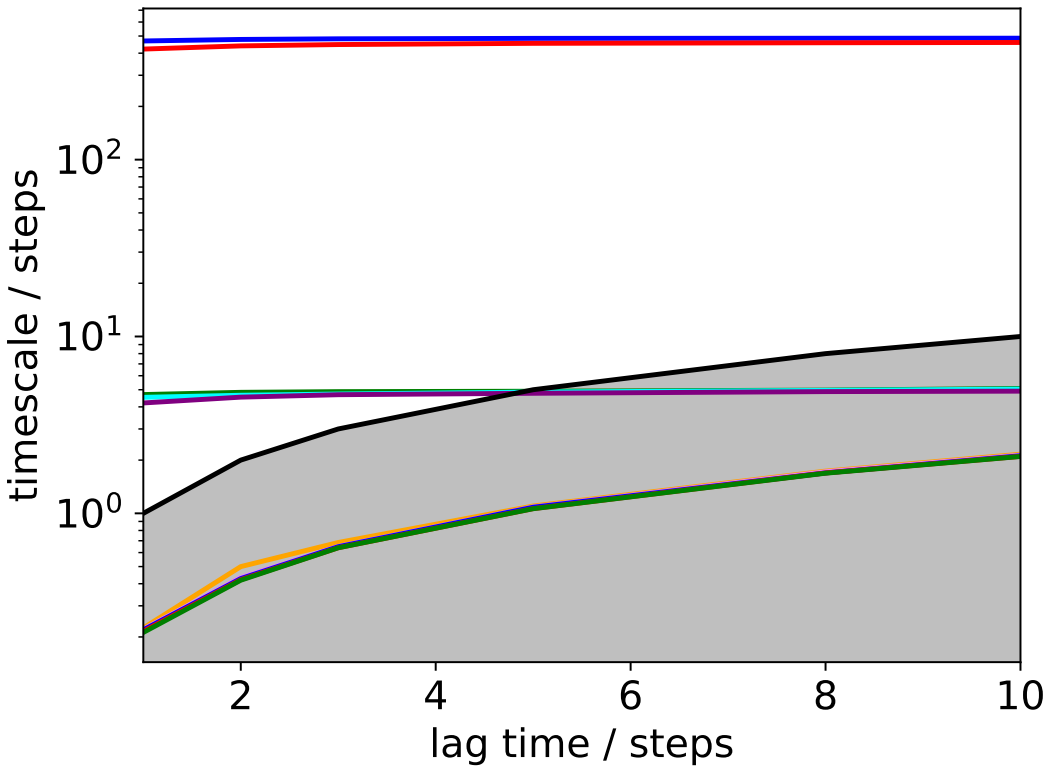

FIGURE 2.8: An example of an Implied Timescale plot. The X axis shows the lag times at which models have been estimated at, while the Y axis shows the timescales of the relaxation processes described by the eigenvalues of the model. Each coloured line indicates a different eigenvalue, with blue being the first and progressing downwards to red etc. The aim is to choose the earliest point on the X axis at which the coloured lines become straight, without processes of interest entering the grey area. The grey area is the region at which the relaxation timescale is shorter than the lag time, and so detail of the processes is lost. In this instance, a Markov time of 3 frames is acceptable.

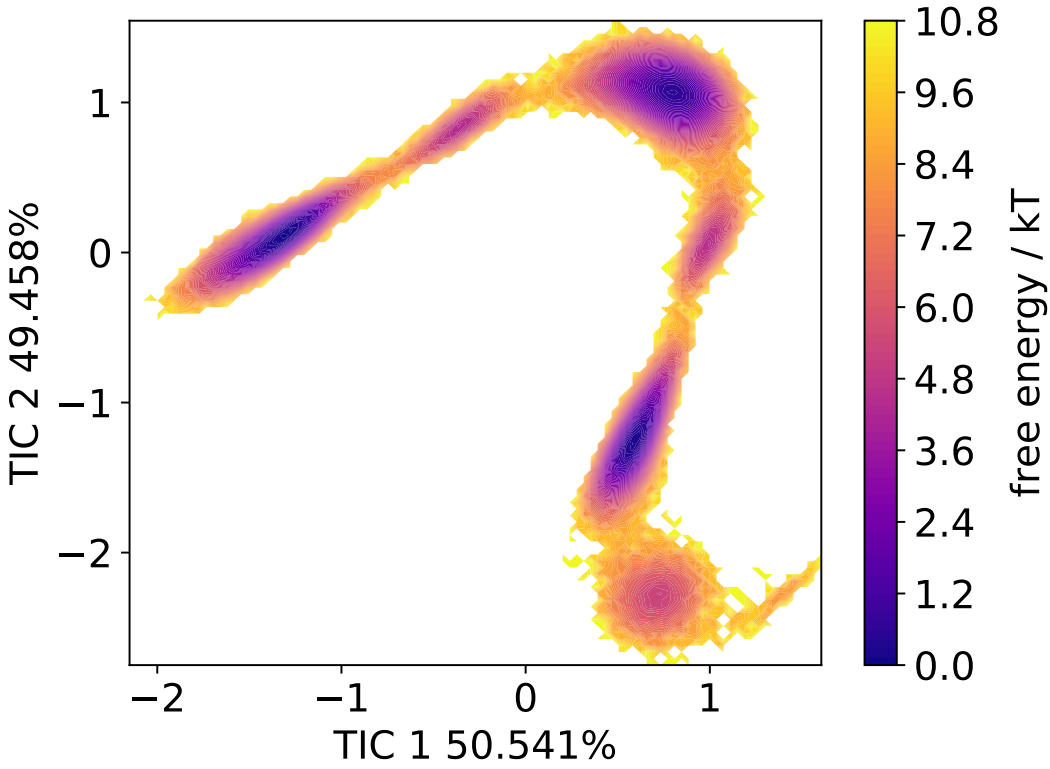

FIGURE 2.9: Example of a free energy diagram.

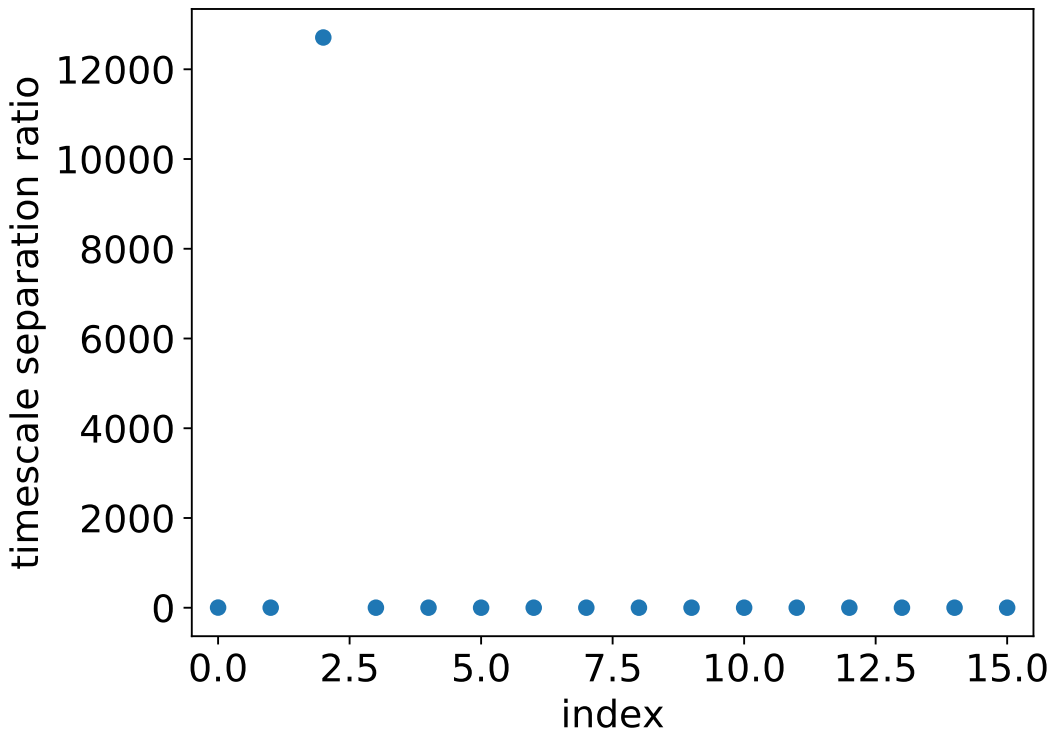

FIGURE 2.10: An example timescale ratio plot to determine the number of metastable states to partition the model into. Adding two to the index indicates how many states the model should be split into, as index 0 refers to the ratio between timescales 1 and 2, index two refers to the ratio between 2 and 3 etc. In this case, the very large value at index 2 indicates 4 states should be used.

3 Diamondoids

3.1 Introduction

Plastic crystals are molecular crystals in which the molecules weakly interact with each other and can reorient themselves within the crystal structure. They were discovered in 1938, but the term plastic crystal was not used until 1948, the term being coined by A. Michils when he observed that certain solid organic compounds were easily deformed [120]. In bulk, these crystals often resemble waxes, and can be easily moulded into shapes. Plastic crystals can be considered as a transition state between "true" crystals and "true" liquids, coming under the heading of soft matter [121].

Diamondoids are a class of hydrocarbons with a diamond-like structure, the simplest of which is adamantane, which take the form of plastic crystals at a variety of temperatures. These hydrocarbons consist of sp^3 hybridised, saturated carbon atoms arranged in cage-like formation, resulting in highly rigid molecular structures. Figure 3.1 shows the structures of the three lower diamondoids, and the diamond-like structure was first determined by Bragg *et al.* in 1913 using X-ray diffraction. As can be seen, each successive diamondoid consists of face-fused adamantane cages, with higher diamondoids often exhibiting many isomeric and structural varations, starting with tetramantane.

Diamondoids occur naturally in petroleum deposits, and can be extracted and purified into crystals of the polymantanes. These diamondoids are then

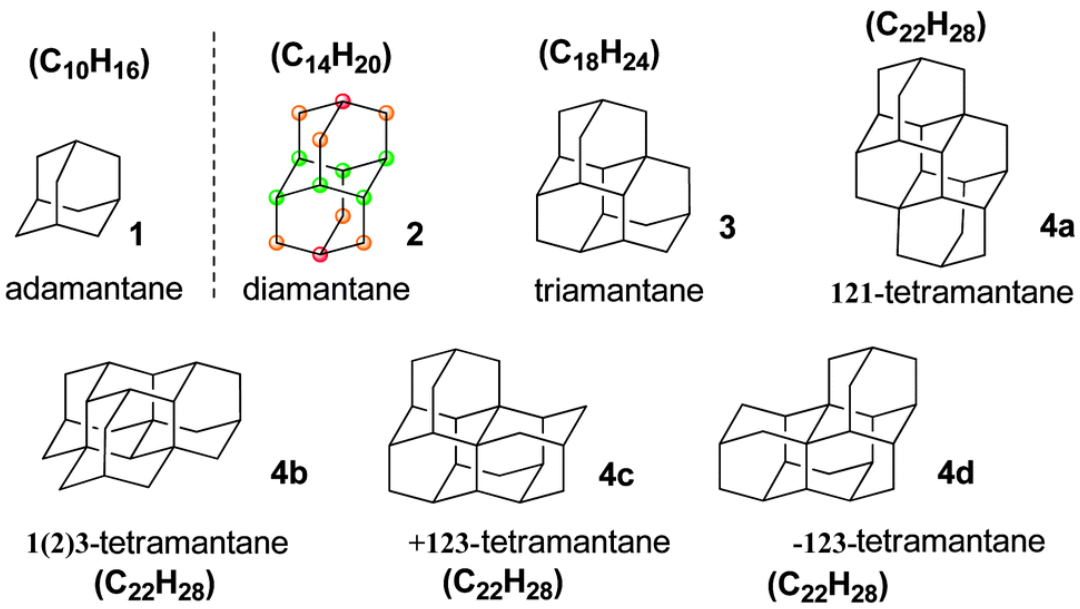

FIGURE 3.1: The lower Diamondoids. Image adapted from [122].

functionalised into a variety of forms, which are then used in applications ranging from the medicinal to the catalytic [122, 123]. Higher and lower diamondoids are of interest in both study and application, with the higher diamondoids being used as molecular approximations of the cubic diamond lattice, albeit ending with C-H bonds. Adamantane, the simplest diamondoid, is often used as a reference sample for solid-state NMR spectra, as its rapid dynamics give sharp lines at a known chemical shift. Adamantane undergoes rapid molecular tumbling, being spatially disordered in its solid state [124], as are many plastic crystals. Numerous studies have been conducted on its structure and disorder, and it is generally well-understood. Diamantane however, is a very different story.

Diamantane is the second of the diamondoids, with 14 sp^3 hybridised carbon atoms and 20 terminating hydrogen atoms. As well as being studied for microelectronics [125, 126], diamantane has been used to improve the thermal stability, chemical stability and solubility of polymers [127–129]. It has a melting point of around 236.5° C and previous studies [130, 131] have shown

the presence of a C_3 rotation axis. The unit cell has four different orientations of diamantane present, as has been shown by XRD studies. While XRD gives us orientations of the molecule within the unit cell, it cannot see the presence of any motion around these orientations. Elucidating these motions and characterising them will give us insight into the dynamics of the molecule, as well as provide a system that allows us to develop a protocol for determining these motions in other, more complex systems.

Plastic crystals' reorientations can take the form of a series of jumps between possible orientations. By rephrasing these orientations as states, we can produce a Markov model using molecular dynamics simulation data. As plastic crystal reorientations tend to be on the order of tens of kilojoules per mole [121, 132, 133], these energy barriers are easily accessible in the course of a simulation, making MD an ideal tool to measure these dynamics.

The diamondoids are present in small quantities in petroleum, and this is currently the only major source of the diamondoids. While diamantane can be synthesised readily [122], triamantane can only be synthesised in small quantities, and the higher diamondoids exhibit major problems during synthesis.

Triamantane consists of three face-fused adamantane cages, with one C_2 axis and two perpendicular mirror planes, belonging to the C_{2v} point group. It was first synthesised in 1966 by Schleyer [134], with a reported melting point of 221-221.5°C. An XRD study of triamantane performed by Carrell and Donohue showed that triamantane crystallised into the Fddd space group, containing 16 molecules in a unit cell, arranged into two symmetrically distinct groups [135]. Cernik reported a gradual phase transition from an ordered phase to a disordered phase starting at 0°C and ending at 30°C. The space group changes above 30°C to Immm, indicating the introduction of pseudo-mirror planes, as shown in figure 3.2 [136].

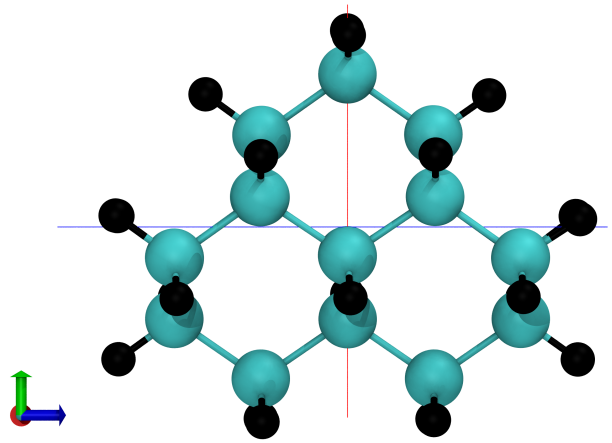

FIGURE 3.2: The molecular structure of triamantane. The red line indicates the C_2 axis and the blue line indicates a pseudo-mirror plane.

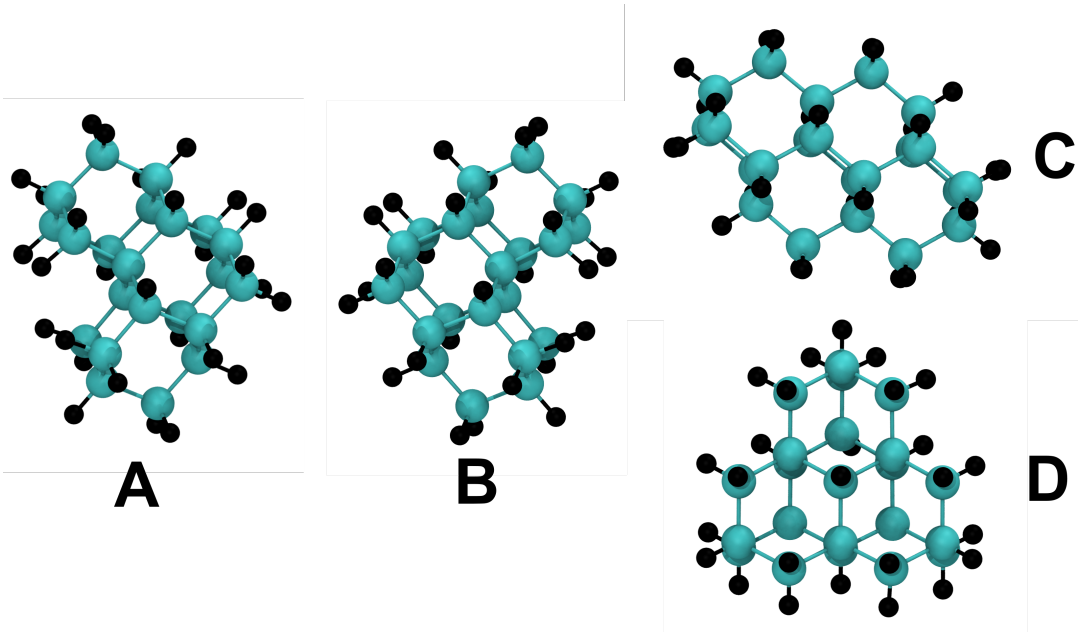

FIGURE 3.3: The three isomers of tetramantane. Isomer A is -123-Tetramantane, B is +123-Tetramantane, C is 121-Tetramantane and D is 1(2)3-Tetramantane. This study focuses on isomer D.

Tetramantane (figure 3.3) is the first of the higher diamondoids, and is also the first of the diamondoids to exhibit geometric isomerism. Three isomers exist, called 123 tetramantane (which is chiral and exists as (+) 123 tetramantane and (-) 123 tetramantane), 1(2)3 tetramantane and 121 tetramantane, all of which are shown in figure 3.3. For this study, we will focus on 1(2)3 tetramantane as structures of 121 tetramantane are unavailable, so a starting point for our simulations is therefore also unavailable. Additionally, 123 tetramantane is chiral, which introduces another layer of complexity into the simulations. While simulations are possible, previous NMR experiments have had the most success in analysing and obtaining data from 1(2)3 tetramantane, and so this isomer was chosen.

Relatively few studies have been undertaken on the dynamics of these diamondoids. A limiting factor may be the lack of available sample, which severely limits the experimental techniques that can be used. MD studies can often be performed starting only with a structure file, however knowing the structure of the crystal requires structural studies on a sample. Once the structure is determined, we can use small amounts of sample to probe the dynamics using a variety of NMR experiments, due to NMR's non-destructive nature. The MD simulations can then be used to further probe the dynamics. In this chapter, I use the methodology outlined in the methods chapter to undertake a detailed study of the dynamics present in the triamantane and tetramantane crystals.

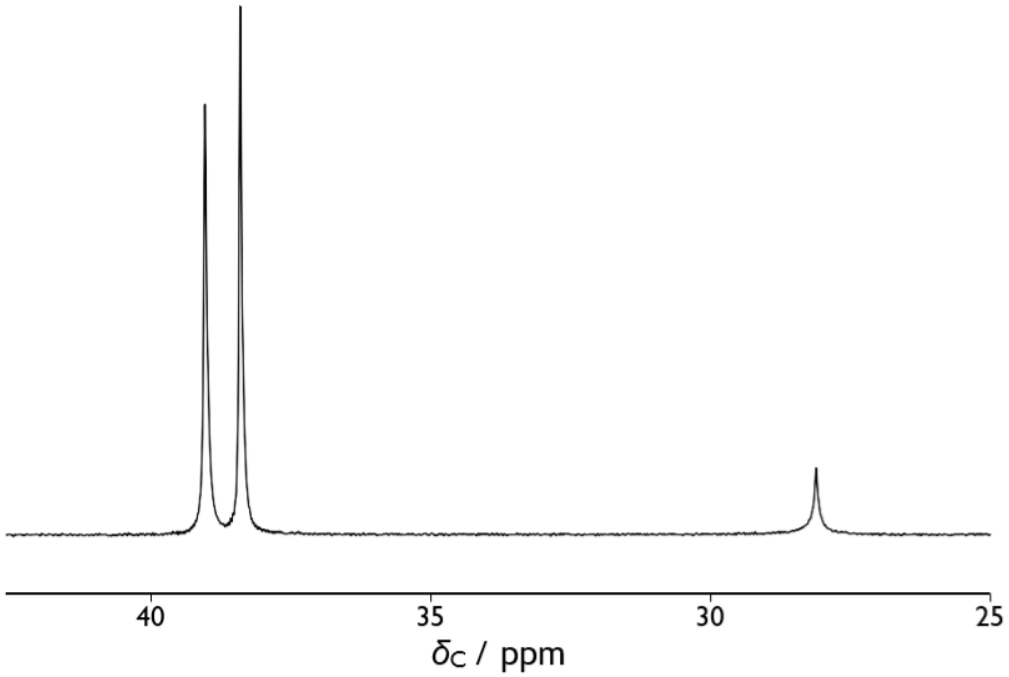

FIGURE 3.4: ^{13}C CP-MAS spectrum of diamantane, supplied by Helen Wickins [137].

3.2 NMR Results

3.2.1 Diamantane

Previous work on diamantane performed within the group focused on characterising the motions present in the system. Diamantane was shown to undergo several pre-melting phase transtions, occuring at 134 °C and 167 °C before melting at 245 °C. The majority of the work focuses on the lower temperature phases, as unit cell parameters or geometric arrangement beyond the phase transitions are difficult to determine due to the extent of the dynamics within the system, and this data is vital for further MD work.

Figure 3.4 shows the ^{13}C CP-NMR spectrum of diamantane. The peaks are reasonably narrow, around 8 Hz, indicating rapid molecular motion. Three carbon environments are present, with the following assignments: 39.1 ppm – C2, 38.4 ppm – C3, 28.1 ppm – C1 [137]. These assignments have been

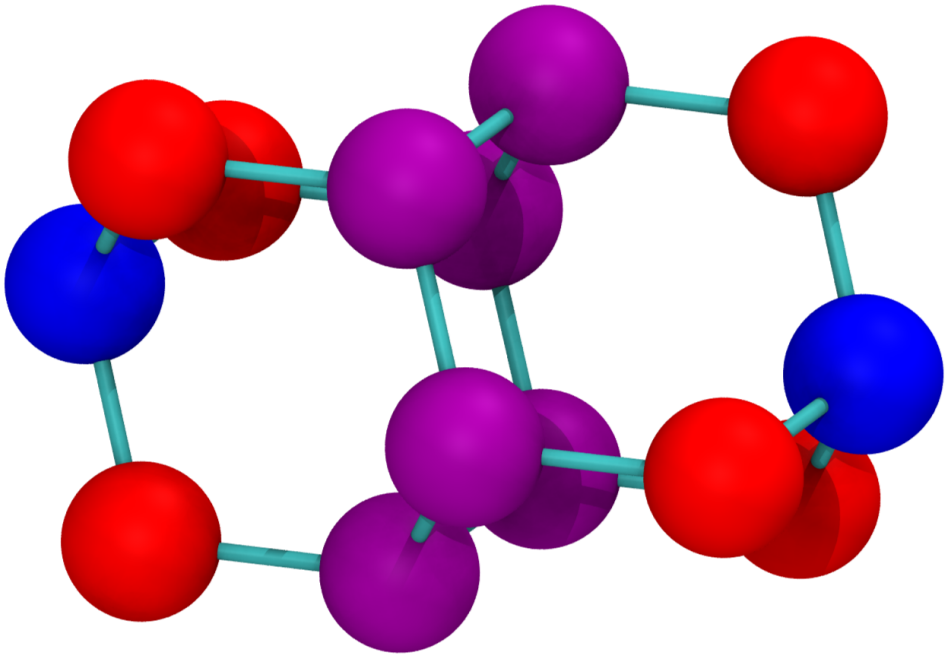

FIGURE 3.5: The structure of diamantane with colour-coded carbon atoms. C1 is in blue, C2 is in red and C3 is in purple. Hydrogens have been omitted for clarity.

shown in figure 3.5.

A series of carbon T_1 relaxation data were obtained, to quantify the proposed motion. Carbon relaxation data is obtained on each carbon site, unlike proton relaxation, which gives molecular averages. This allows us to infer useful information about the nature of the motion, depending on which signals decay faster/slower.

Figure 3.6 shows the carbon T_1 values recorded for diamantane, as taken by previous group members [137]. The CH_2 T_1 values are shorter than the CH values, while the T_1 of C1 is much higher than the values for C2 and C3. This provides good evidence for the presence of the C3 rotation, as the C1 carbons will be aligned along the rotation axis, increasing their T_1 relative to the more mobile C2 and C3 carbons, as the C_3 rotation does not affect the orientation of C-H bond for the C1 carbon atoms. These relaxation times were fitted

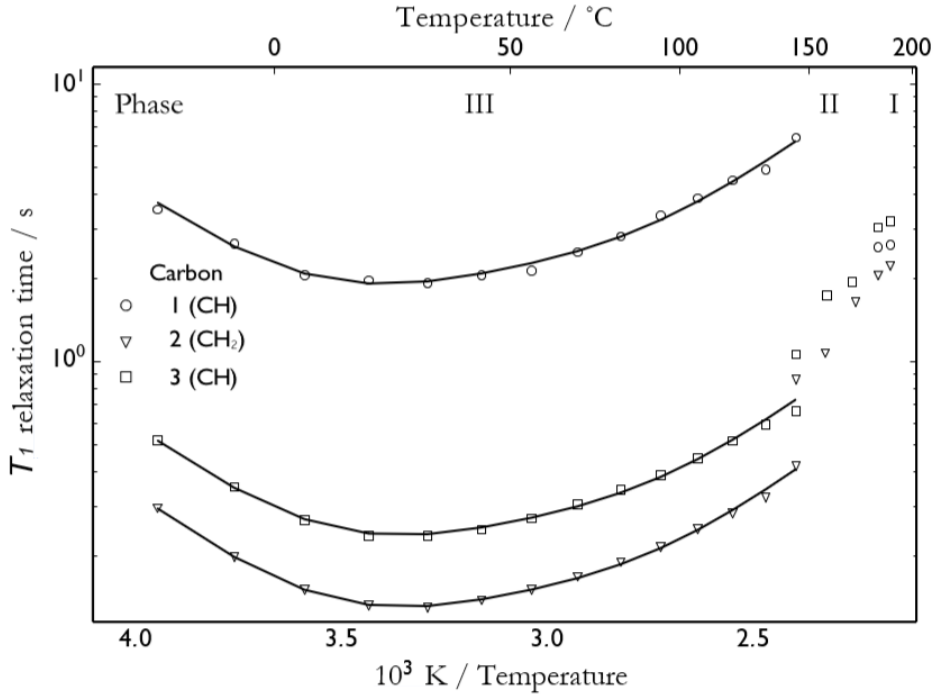

FIGURE 3.6: ^{13}C T_1 relaxation data for diamantane. The solid lines show the result of the Arrhenius-fitting of the data, as described in the text. Data supplied by Helen Wickins [137]

using an in-house program between -20 to 140° C, obtaining an activation energy of 17.4 $\pm$ 0.4 kJ mol^{-1} as average for rotation between the three sites. We assume an Arrhenius-type model for the temperature dependence of the relaxation times.

3.2.2 Triamantane

Preliminary NMR results for triamantane have shown that recording an NMR spectrum at temperatures over 30 °C is extremely difficult due to significant line broadening. This broadening was theorised to be a result of molecular motion approaching the rate of ^{1}H decoupling, which produces an interference effect with the spectrum [138]. This may be the same motion described by Cernik [136] as an order-disorder phase transition, characterised by rotating around a pseudo-C_2 axis or mirror plane. At around 30 °C, the molecular

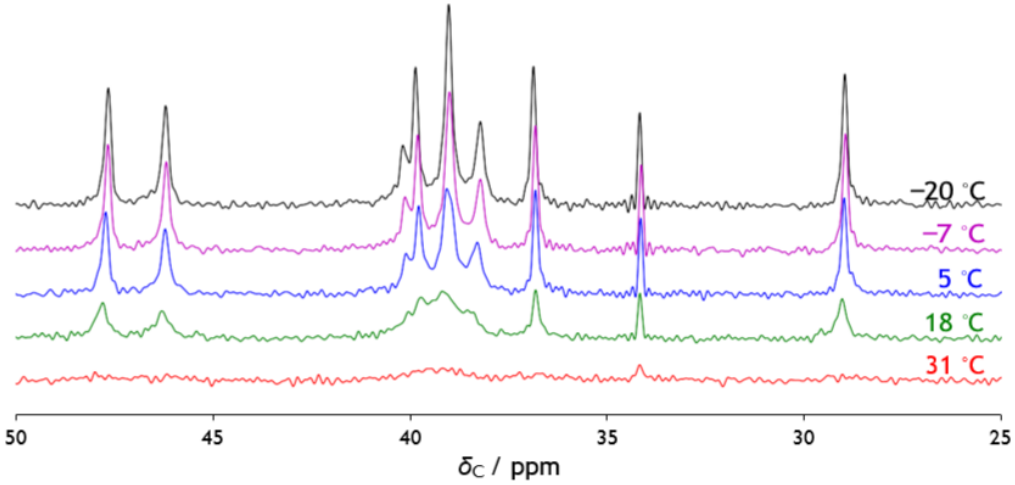

FIGURE 3.7: ^{13}C CP spectra of triamantane over a range of temperatures. As the temperature increases, the line broadening increases significantly, as a result of increased molecular motion in the order of the ^{1}H decoupling. Spectra supplied by Helen Wickins [137].

rate of motion is around 50 kHz, which corresponds to an average transition time of around 20 microseconds. Additionally, a phase transition occurs between 151 °C and 164 °C, which means performing MD simulations at these temperatures becomes impossible, as the 1st order phase transition results in an unrelated unit cell, which is extremely difficult to correct for during the simulation. This transition can be seen in figure 3.6, as the relaxation values increase rapidly and no longer obey an Arrhenius type pattern.

3.2.3 1(2)3-Tetramantane

Preliminary results for tetramantane show similar results as for triamantane: line broadening occurs at around 12 °C, again as a result of molecular motion matching the rate of ^{1}H decoupling, producing an interference effect. Above 40 °C, this motion becomes faster, and ceases to interfere with the recording of spectra, and at 163 °C, line widths from a direct excitation experiment becomes very narrow, indicating rapid molecular motion. The relaxation data shows that hydrogens pointing parallel to the C_3 axis are significantly slower

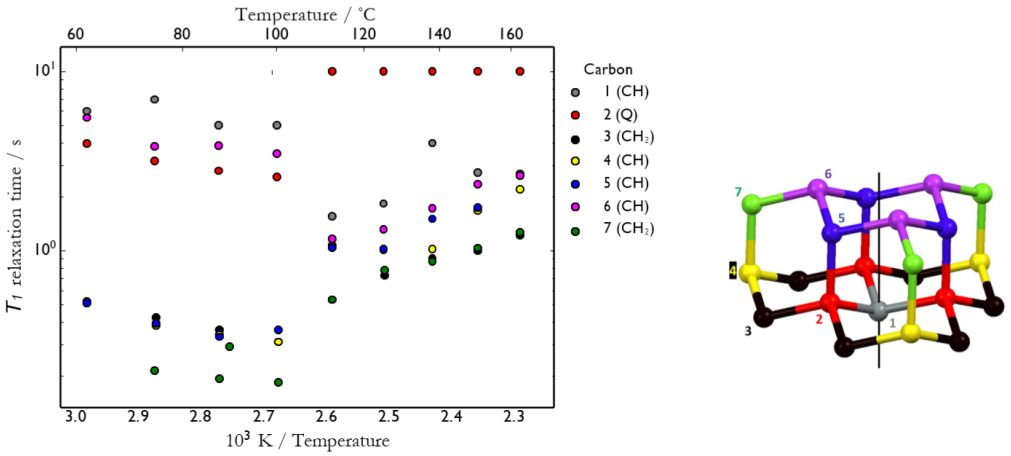

Figure 51: Carbon T_1 relaxation measurements.

FIGURE 3.8: The preliminary relaxation data for Tetramantane. At higher temperatures, carbon atoms bonded to more hydrogen atoms are the fastest to relax, which is typical of isotropic motion. Below 110 °C, the relaxation times change so that carbon atoms with hydrogens pointing along or parallel to the supposed C_3 rotation axis relax slower. This is good evidence of a C_3 rotation. Figure adapted from reference [137].

to relax than other protons, indicating a motion around the C_3 axis, which gives us a good set of data to begin our analysis with [137].

3.3 Method Development

3.3.1 Diamantane

Atomistic molecular dynamics simulations were performed on a system of 256 diamantane molecules using periodic boundary conditions in the xyz directions. The initial atomic positions were taken from a CIF file obtained from the CCDC (Refcode: CONGRS) and the unit cell copied in each direction until an appropriately sized supercell was obtained. The supercell was deemed large enough when half of the smallest edge (4.05 nm) was longer than the electrostatic interactions cut-off distance (1.4 nm), as determined by the force field .

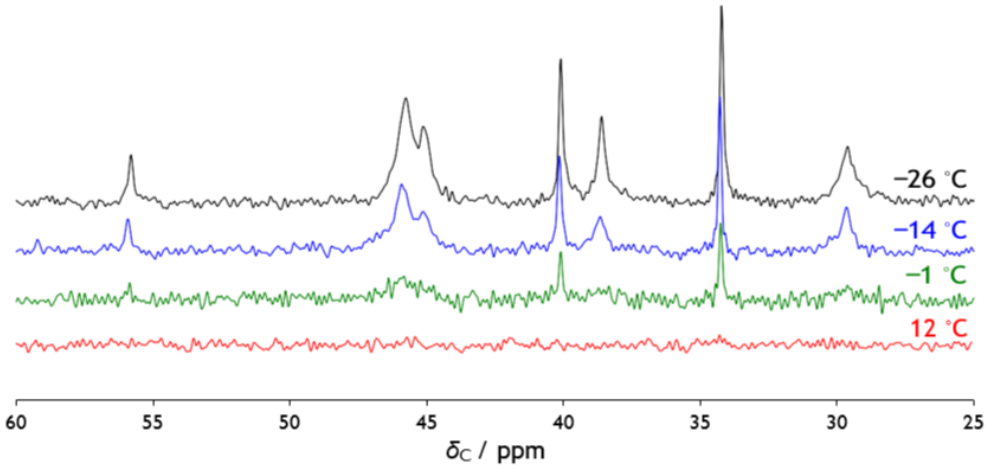

FIGURE 3.9: ^{13}C CP spectra of 1(2)3-tetramantane over a range of temperatures. This shows similar behaviour to the triamantane, with the broadening occuring at lower temperatures indicating slower overall motions. Spectra supplied by Helen Wickins [137]

The starting atomic coordinates in the CIF file also acted as the basis for the force field. A single diamantane molecule was extracted from this file, and the topology for this molecule assigned using the CHARMM AA force field [97]. The CGenFF tool [98, 139] was used to create this forcefield, as discussed in the methodology section. The penalty score calculated from the assignment was very low (< 5), and thus the assignment was used as provided.

The simulations were run using the GROMACs 2016.4 simulation package [105–107, 111]. Long range electrostatics were calculated using a cut-off of 1.2 nm. The configuration started at 300 K, and was equilibrated for 1 ns. An annealing run was then performed, raising the temperature from 300 K to 350 K over a period of 100 ns, increasing the temperature by 10 K every 20 ns. During the annealing runs, the temperature of the system is steadily changed, increasing or decreasing the amount of energy available to the system. This takes the form of an increase or decrease in kinetic energy, which balances with the potential energy after some time. Letting the system run at

a steady temperature after annealing allows the two energies to equilibrate, and once they are both stable, the system is converged and ready to undergo a production run. Snapshots of the system were extracted at 310 K, 325 K and 350 K, and were equilibrated for an additional 10 ns before being used for the production runs. An additional annealing run was performed to obtain a starting structure for a lower temperature, as each annealing run can only be used to raise or lower the temperature. This run lowered the temperature from 300 K to 250 K over 100 ns, decreasing the temperature by 10 K every 10 ns. A final snapshot at 250 K was equilibrated for 10 ns before being used for a production run.

Each temperature run was sampled every 5 ps for 200 ns using the Parrinello-Rahman thermostat set at the appropriate temperature and the Nosé-Hoover barostat set at 1 bar. An anisotropic pressure coupling was used to allow the system some flexibility in its expansion due to changing temperatures. Using this pressure coupling worked for the majority of the simulations, however upon finishing the 300 K to 250 K annealing step, the 250 K simulation resulted in the simulation box twisting out of shape, despite prolonging the equilibration time. To remedy this, the anisotropic pressure coupling was still used, but the off-diagonal components were set to 0, ensuring the box shape stayed rectangular. As we know the crystal structure is cubic at this temperature range, this is a reasonable adjustment to make. The end point of the 300 K to 250 K simulation was used, as the box remained cubic during this, with the additional restraints added in before the further simulations.

The analysis of the simulations utilised a mixture of custom-written code and the PyEMMA 3.5.4 module [116], as detailed in the methods chapter. A system-specific summary of the analysis follows.

Initial attempts at analysing the simulations focused on the positions of key atoms, theorising that these positions would hold the orientation of the molecule

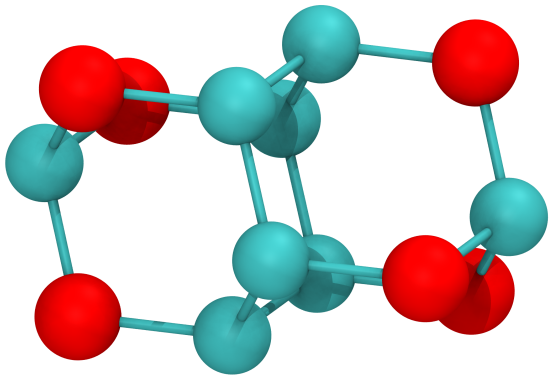

FIGURE 3.10: The structure of diamantane with the carbon atoms initially chosen for the analysis highlighted in red. Hydrogen atoms have been omitted for clarity.

at large. These atoms have been highlighted in figure 3.10. A script written in TCL, the programming language available in the visualisation software VMD, was utilised in this case, as the script can integrate directly with the trajectories loaded in VMD, and makes use of VMD's atom selection tool. The script directly extracted the x, y and z of the carbon atoms in each diamantane molecule. In order to offset their translational differences, the script placed each molecule's centre of mass at the origin (co-ordinates 0,0,0) before extracting the co-ordinates. In this way, any major translational difference is removed, in theory leaving just the orientational differences.

Analysis of these extracted co-ordinates proved inefficient. As we were supplying three sets of co-ordinates (x, y and z) for each chosen atom for each molecule into the system, we were obtaining many principal components in return, however the vast majority of them contributed neglibly to the description of the motion. This lack of contribution indicates a large degree of redundance between the components, as each subsequent component describes a new direction of variance. Components with small contributions

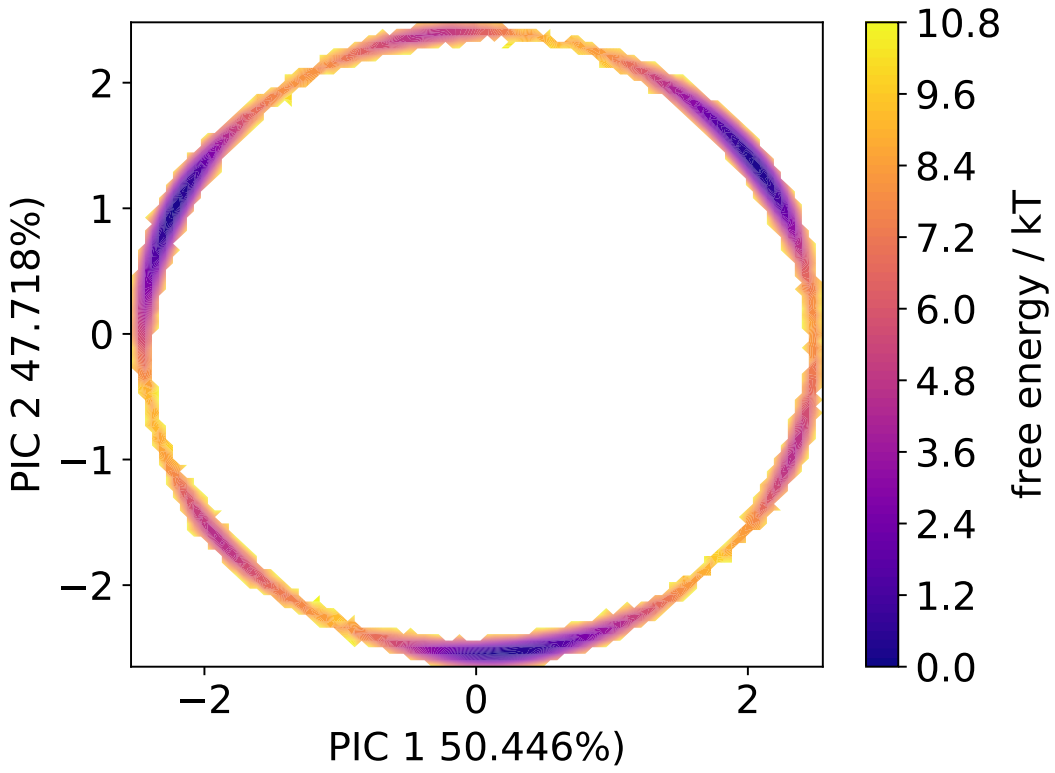

FIGURE 3.11: Free energy graph derived from the first two principal components of the diamantane positions. It is expected that both PICs should have a contribution of around 50% each, however PIC 2 has a lower contribution than this. This is because the C_3 motion is not the only motion occurring in the system, as some molecular "wobbling" is to be expected in the simulations.

indicate that the motions described by that component are describing very small variances in the data. Figure 3.11 shows an example of running these co-ordinates through the PCA analysis. As can be seen, there is a thin ring of explored coordinates around the edges, with the majority of the space left blank. It is possible to see the locations of the three expected minima, and we can make an educated guess as to the process that relates these minima. However, as we are putting coordinates in, it is difficult to determine what the components mostly relate to. Additionally, processing this many co-ordinates takes an extended amount of time, and a more computationally efficient description of the motion can be obtained.

A different strategy was employed to analyse these co-ordinates. The positions were first converted into two vectors, which describe the overall orientation of the molecule with respect to the box axis. They were chosen as the C2-C9 vector, and the C12 to halfway between C5 and C7 vector, as shown in Fig 3.12. The ϕ and θ polar co-ordinates for these vectors were extracted across the course of the trajectory and used as the basis for the PCA/TICA and MSM modelling input. The results of the previous extraction script were thus reprocessed to convert them into spherical vectors, discarding the magnitude of each vector. These equations are detailed as equations 3.1 and 3.2.

$$\phi = \text{atan2}(\frac{y}{x}) \tag{3.1}$$

$$\theta = \arccos(\frac{z}{r}) \tag{3.2}$$

Where x, y and z are the cartesian components of the vector, θ is the inclination of the vector and ϕ is the azimuth, the angle going clockwise from the x plane. This provides us with four values overall for each molecule. This is clearly a huge decrease in the amount of data, yet still contains the information of relevance.

An additional complication comes from the four diamantane molecules present in the unit cell. Each diamantane is related to the others in the cell by a 90°rotation, so the vectors for each orientation will take very different values from each other. While this is useful for observing transitions between these orientations, it could obscure the individual C_3 rotations we expect to see. To correct for this, the different orientations were analysed both separately and together, to allow us to see the rotations within each orientation and any potential transitions between them.

The PyEMMA 2.5.4 Python module was used to analyse the ϕ and θ angles. A set of angles was generated for each molecule as described above, and

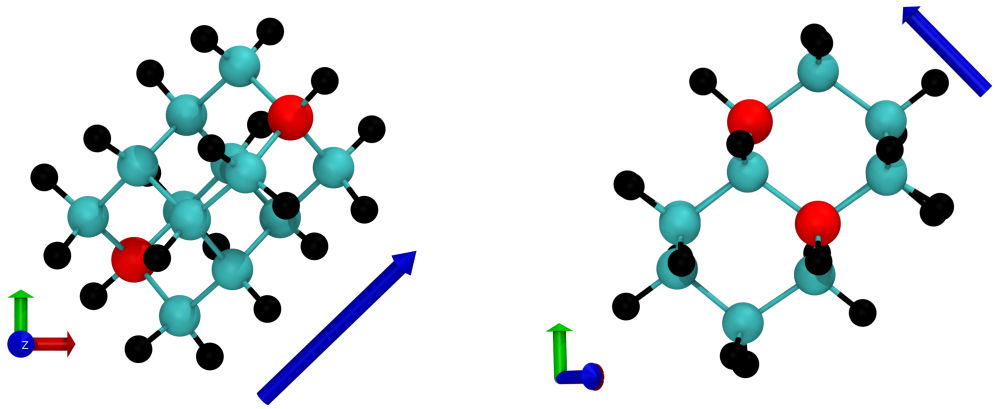

FIGURE 3.12: The analysis vectors chosen for diamatane.

each of these treated as a separate trajectory by the module. The group of reduced trajectories was then put through PCA analysis, as detailed in the methods chapter of this work. The resulting PICs were used as the basis for the Markov state modelling. The data first requires clustering, using the k-means clustering method described in the methodology. A range of cluster centres were used to determine a suitable amount that reasonably described the entire data set.

The number of cluster centres directly affects the accuracy of the MSM. As can be seen in Fig 3.13, using only a few cluster centres causes an inaccurate description of energy landscape to be created. Increasing the number of cluster centres allows us to minimise this error but results in a greater computational cost. Striking the balance between accuracy of the model and computational cost is required when creating these models.

During the creation of the model, we create a chart of the implied relaxation timescales of the model. These timescales are affected by the number of cluster centres, most notably in that they become smoother and converge earlier as we increase the number of centres. We produced implied timescale charts and compared them to determine when they converged, as shown in figure

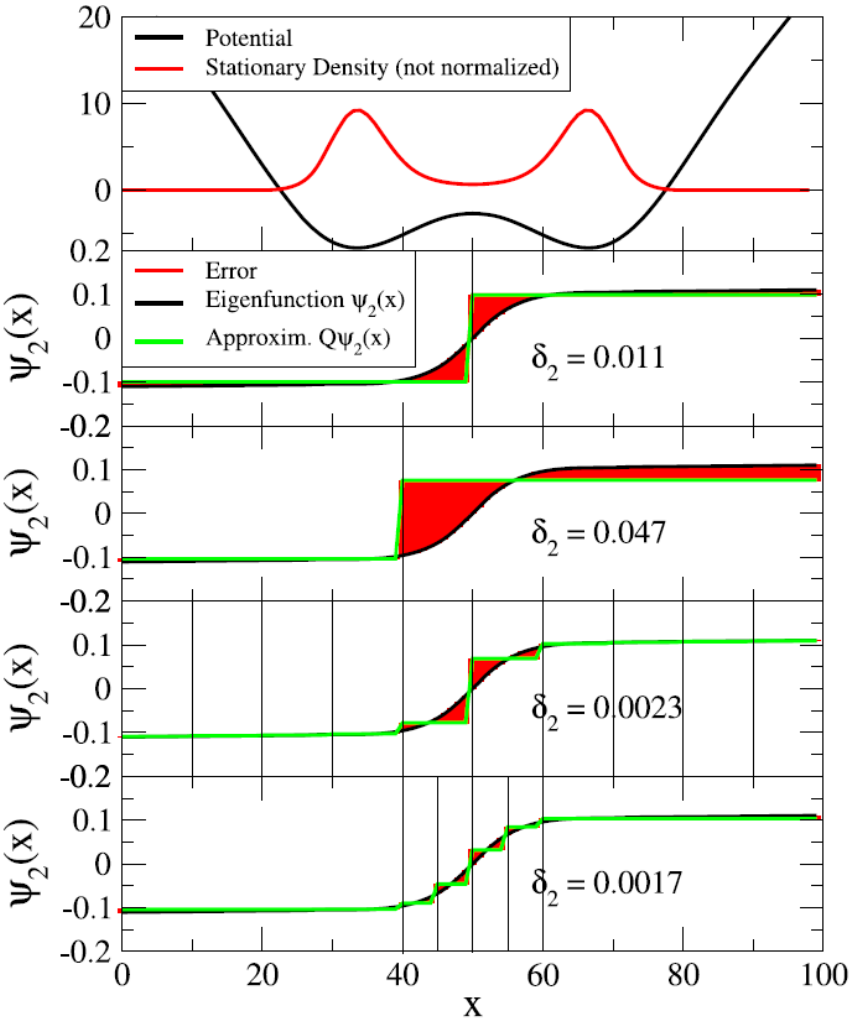

FIGURE 3.13: Illustration of the error associated with insufficient clustering. The upper box shows an example energy landscape, while subsequent boxes show various levels of clustering. The region in red is the error associated with each clustering method. Focusing clustering around a known transition point (a point along the free energy surface that separates one state from another) lowers the error, but increasing the number of clusters overall achieves the same effect. Determining transition points is difficult, if not impossible, to do before constructing Markov models from the data, and so high levels of clustering are used to minimise the error. This figure was taken from reference [140].

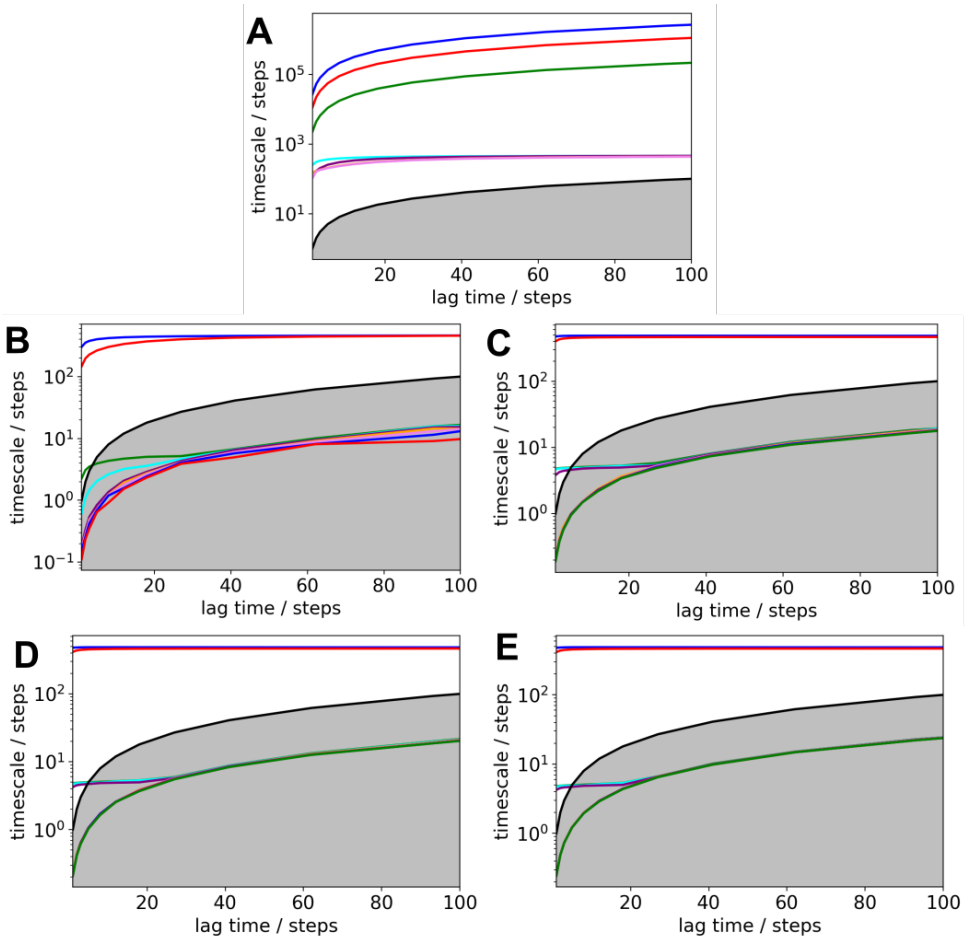

FIGURE 3.14: Implied timescales after clustering the diamantane system with varying numbers of clusters. The plots refer to clustering with: A) 8 B) 32 C) 64 D) 128 and E) 256 cluster centres. The grey region represents the area where the timescale is shorter than the lag time of the Markov model, indicating that choosing that lag time results in a model that cannot accurately observe the timescale of interest.

3.14. From the figure, we can see the timescales converge much earlier as we increase the number of cluster centres, although there is no obvious change between 256 and 512 centres. As such, 256 cluster centres were used, as this preserves the accuracy of the model while speeding up the clustering step.

During the clustering process, the cluster centres are assigned as distinct states of the molecule during the trajectory. Once the states have been identified, a Markov state model can be produced. In brief, we count how many transitions there are between each pair of states a and b at a given lag time τ,

i.e. if the system is in state a at time t, how many times did a transition to state b at time $t + \tau$ occur. Once the Markov model has been produced, the results are coarse-grained using PCCA to create a hidden Markov model, allowing us to see the metastable states within the system (see section 2.5.3 and 2.5.4). A key portion of the analysis is identifying the particular orientation of the molecule in each of these metastable states, and the pathways the molecules take when going from state to state. A Python script was used to extract the timeframe and molecule number of examples of each of the metastable states, and a Tcl script used to extract snapshots of the molecule at the right time frame. Multiple examples of each metastable state were extracted, to obtain an average idea of the clustered state.

For both systems, the same overall simulation protocol was used, with minor differences in specifics. Starting CIF files were obtained from the CSD [104] for triamantane (refcode: TRIAMT01) at 293 K and in-house for tetramantane at 120 K. A topology for each molecule was generated using the CGenFF [97] tool. A low overall penalty score was found for each of the systems, with triamantane obtaining a score of 1.2 for bonding parameters and 0.4 for charge, and tetramantane obtaining a score of 1.2 for bonding parameters and 1.3 for charge. Therefore, the topologies were used as is. Following this, each system went through two equilibration steps for a total of 15 ns of equilibration. These steps were used to stabilise the system from the starting configuration. After this, both systems went through an annealling simulation to raise the temperature to desired simulation levels.

For triamantane, the simulation box was 3.86 by 5.41 by 6.61 nm along the x, y and z axis respectively. The system started at 293 K and the temperature raised to 310 K after 30 ns. After this, the system went through a cycle of equilibration and temperature raising, with simulation length of 10 ns for each equilibration and a length of 30 ns for each increase in temperature. The

temperature was raised from 310 K to 340 K, 350 K, 380 K and finally 400 K, equilibrating after each raise. Snapshots were extracted at 310 K, 350 K and 400 K for further simulation.

For tetramantane, the simulation box was 3.08 by 3.59 by 4.46 nm along the x, y and z axis respectively. A similar protocol to that used for triamantane was followed, with 30 ns raising periods followed by 10 ns equilibration periods. As the starting CIF file was provided at 120 K, the temperature increments for this were: 120 K, 150 K, 180 K, 210 K, 240 K, 270 K, 300 K, 310 K, 340 K, 350 K, 380 K and 400 K. Snapshots were taken at the end of the equilibration periods for 310 K, 350 K and 400 K for further simulation.

The production simulations were run for 400 ns total, with a timestep of 1 fs, sampling the coordinates every 5 ps. A long simulation and high sampling rate is required, as the motion in these systems is thought to be in the 10s - 100s of ns timeframe, however there may be rapid motion that the NMR does not see (such as bond vibrations), so a high sampling rate is needed.

While the general methodology has been discussed extensively in Chapter 2, there are several key parameters that need to be adjusted for the analysis depending on the system. These parameters include the TICA lag time, the MSM lag time, the number of cluster centres and the number of metastable states. Detailed discussion on each of these points has been provided in Chapter 2, so here I will present the optimisation performed for each system, along with the reasoning for each selection.

3.3.2 Triamantane

During the simulation protocol outlined above, a major issue arose with the triamantane system. The box size of system fluctuated with temperature as expected, but upon running long runs at a stable temperature, the box would often change shape, leading to a steady "wobble" in the trajectory. The box

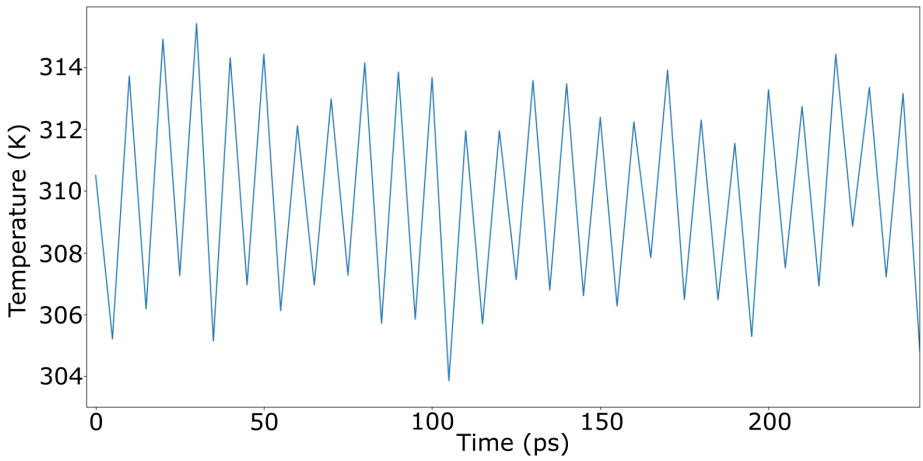

FIGURE 3.15: This graph shows the fluctuation in temperature over time during the first 250 ps of the simulation. The time difference between peaks is around 10 ps, and so this was chosen as the temperature coupling time.

size should be stable at a constant temperature, and so this problem needed to be addressed. The pressure and temperature coupling times were investigated first, as if these are not long enough to encompass the natural fluctuations in pressure or temperature, this can have a knock-on effect that extends the box size. Choosing suitable levels involves a simple process: the simulation is run for a few hundred picoseconds, and the pressure and temperature plotted over this time. Measuring the distance from peak to peak of the resulting graph gives an estimate of appropriate values to put into further simulations. In this case, as can be seen from figures 3.16 and 3.15, a temperature coupling time of 10 ps is required and a pressure coupling time of 50 ps is required. After putting these values in, the box still wobbled, requiring further investigation.

Changing the thermostat and barostat types could also affect the box size. The production runs were originally simulated using the Parrinello-Rahman barostat and Nosé-Hoover thermostat, as these produce high accuracy simulations. However, if the system is not well-equilibrated, these algorithms

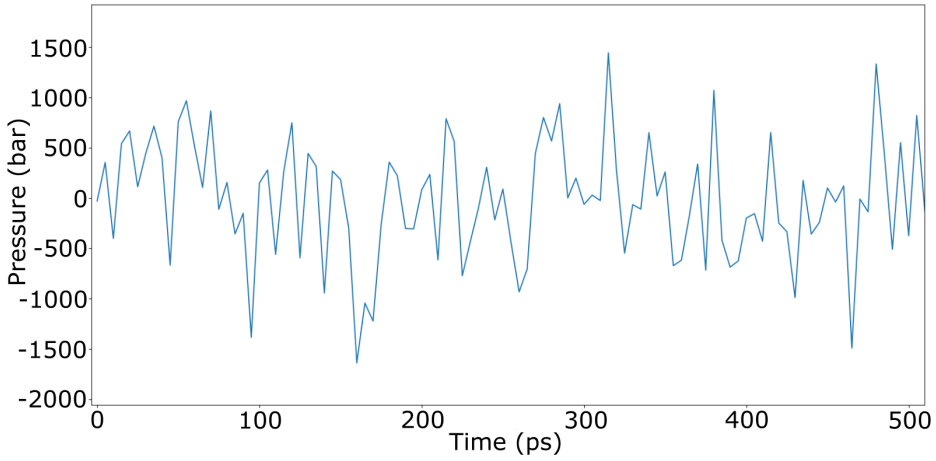

FIGURE 3.16: This graph shows the fluctuation in pressure over time during the first 500 ps of the simulation. This graph is much noisier than figure 3.15, but there are repeated dips in the graph roughly every 50 ps. As a result, 50 ps was chosen as the pressure coupling time.

can be both time-consuming and lead to irregularites, such as the wobble observed. The temperature raising runs were performed using the V-rescale thermostat and the Berendsen barostat, both of which are less accurate but more tolerant to deviations from the equilibrium. The temperature raising runs did not show the wobble, and as the choice of algorithm was one of the key differences, it was thought that this could be causing the problem. However, performing production runs using these algorithms would lead to less accurate results, so a different solution was sought.

Another parameter in the simulations is compressibility. We had set the pressure coupling type to anisotropic, allowing the box to change in size along any axis as it increases in temperature. While this was useful while performing the temperature ramps, as the box could expand as needed, when running at a constant temperature, the box should stay a constant size. Using the same method as for the diamantane case, we set the xy/yz, xz/zx and yz/zy compressibility values to 0 to ensure a rectangular box is preserved. As we know from the crystallography that the box is rectangular, making this

change allows us to use the higher accuracy thermo- and baro-stats, while eliminating any wobble within the box.

Upon performing these tests, the parameters for the run were as follows: The Nosé-Hoover thermostat was used, with a time constant of 10 ps. The Parrinello-Rahman barostat was used, with a time constant of 50 ps, and compressibilities of $4.5x10^{-5}$ 1/bar for the first three components, and compressibilities of 0 for the last three, to ensure a rectangular box shape.

Entering the entire set of coordinates for triamantane into the PyEMMA module would result in overly long processing times and inaccurate data. The inaccuracy arises from the fact that each triamantane has its own unique position in the simulation box, and simply analysing the raw coordinates into something common between all molecules is not possible using TICA or PCA alone. To aid this, each molecule can be translated to the center of the box, recording the new sets of coordinates across the entire trajectory, allowing all the molecules to share the same space. However, there are still 42 atoms, each with 3 coordinates, meaning 126 coordinates total for each step in the trajectory, of which there are approximately 40000. There are also 432 triamantane molecules in the system, so the sheer size of data would take an inconveniently large amount of time to process.

To address this problem, we can use the same method we used for diamantane, extracting vectors to represent the entire molecule. To obtain the vectors, the coordinates of four carbon atoms, detailed in figure 3.17, were extracted and converted into vectors, which together describe the absolute orientation of the molecule. Reducing the data down in this fashion allows us to make the data sets far more manageable, while retaining the pertinent information. These cartesian vectors were then converted into spherical polar co-ordinates, using equations 3.1 and 3.2.

The TICA lag time is the first parameter to be optimised. Using the 310 K

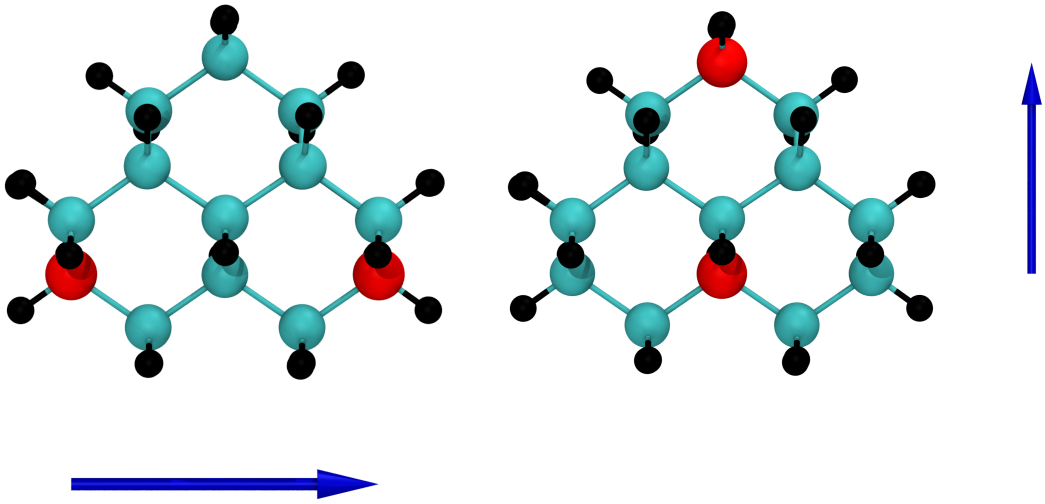

FIGURE 3.17: The vectors selected for analysing triamantane. The red atoms in each diagram represent the start and end of the intramolecular vectors extracted to perform the analysis upon. The atom numbers are: C6 (left) and C6A (right) for the first diagram and C1 (bottom) and C8 (top) for the second diagram. The atom numbering is obtained from the starting CIF file.

system and a series of lag times (1, 5, 10, 50, 100, 500), the data underwent the TICA transformation, was clustered with 128 cluster centres and implied timescale plots created. The resulting plots are presented in figure 3.18. The plots presented represent lag times of 1, 10 and 100. From the plots, there is not much difference between each of the lag times, and so a lag time of 10 simulation steps was chosen for the analysis. Changing the TICA lag time increases the size of the "window" upon which analysis is performed, and the lack of change indicates that the motion seen at these temperatures is extremely fast, as shown by the timescales appearing in the grey area, and are therefore unlikely to be fully resolved within the Markov model.

Once a suitable TICA lag time has been chosen, a reasonable number of cluster centres needs to be chosen. A range of cluster centres (16, 32, 64, 128, 256, 512) was used and the data was clustered accordingly. Implied timescale plots were then created based upon the clustering, and have been presented

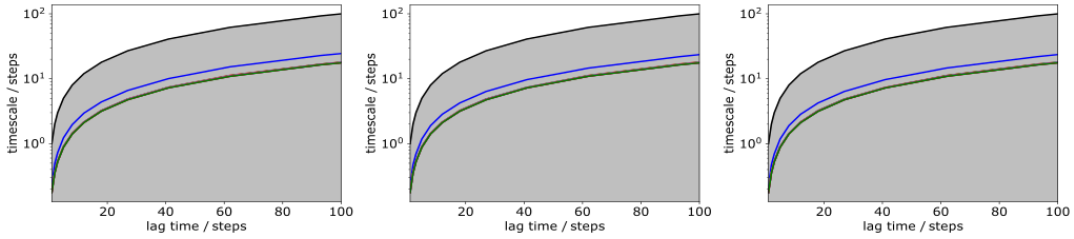

FIGURE 3.18: Implied timescale plots for triamantane at a range of TICA lag times for the 310 K simulations. From left to right, the lag times are 1, 10 and 100 simulation steps, with each step lasting 0.5 fs, and 256 cluster centres used. The grey region indicates where the length of the timescale is smaller than the lag time the model was generated for. Important to note is that the lag time along the x axis is the lag time for generating the Markov matrix, which is distinct from the TICA lag time, which is used to calculate the autocorrelation of the data.

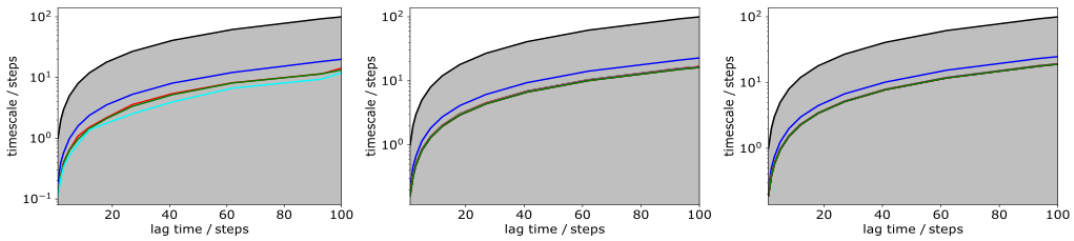

FIGURE 3.19: Implied timescale plots for triamantane at a range of cluster values for the 310 K simulations. From left to right, the number of cluster centres are 8, 64 and 256. The data first underwent a TICA transformation using a lag time of 10. Each step lasts 0.5 fs.

in figure 3.19.

These plots represent cluster numbers of 8, 64 and 256. At 8 cluster centres, we see that rather than a smooth curve that levels out, there is a small uptick in the timescales as we reach the 100 lag time mark. This indicates that the timescales of motion start to rapidly increase at this point. Creating an accurate MSM requires that increasing the lag time should see no significant effect on the timescales of the motion, and so using 8 clusters is inappropriate. 64 and 256 clusters do not show this uptick, but rather show a linear

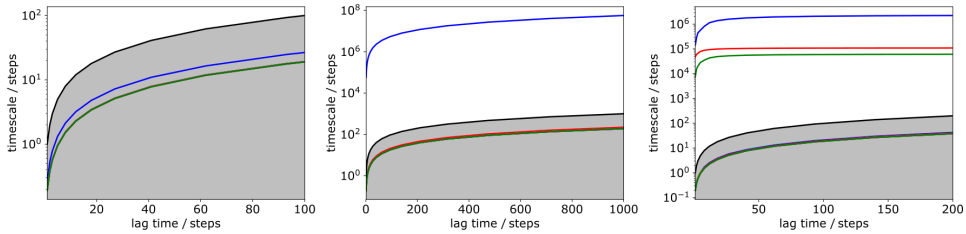

FIGURE 3.20: The implied timescale plots for Triamantane. From left to right, the plots are for 310 K, 350 K and 400 K. These were generated using a TICA lag time of 10 steps and 256 cluster centres.

relationship between the timescales and lagtime. This indicates that the motions described by the lines are not adequately converged. However, producing similar graphs for the higher temperatures does show clear straight lines showing suitable lag times to be used to create the MSM, and these graphs have been detailed below.

Choosing the correct MSM lag time is the next step in the analysis. The MSM lag time is distinct from the TICA lag time, and specifies the window size for calculating the MSM, as described in section 2.5.2. This may be unique for each temperature, and so implied timescale plots were created for each temperature, and have been represented in figure 3.20.

While the longest timescale (indicated by the blue line) steadily increases beyond the edge of the graph for both 310 K and 350 K, this stabilises at 400 K. Looking at the next two longest timescales, we can see that by around 100 simulation steps both lines have stabilised for the 400 K system, whereas the other temperatures show no such behaviour. As such, 100 steps was chosen as the MSM lag time for all three systems.

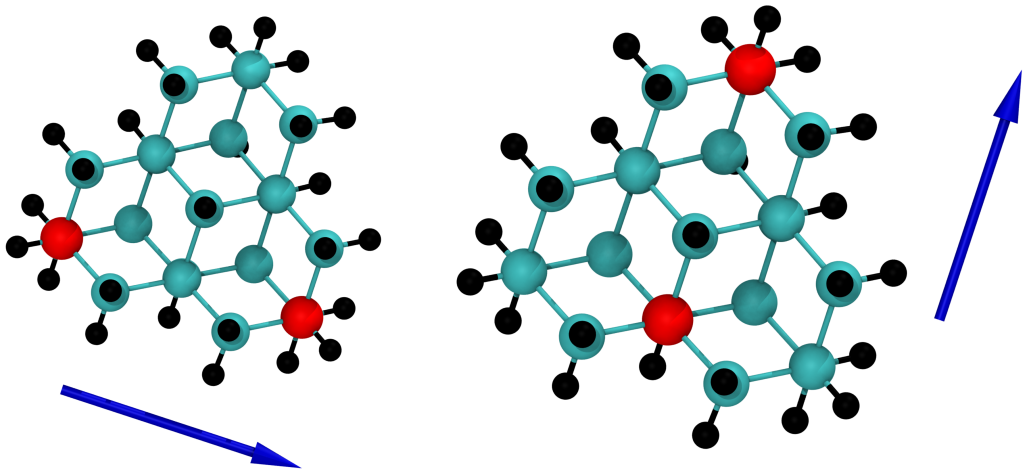

FIGURE 3.21: The vectors selected for analysing tetramantane.

3.3.3 Tetramantane

With the previous diamondoids, the use of intramolecular vectors rather than Cartesian co-ordinates was used to reduce the amount of data requiring analysis, speeding up the process, while also retaining the information of interest (the overall orientation of the molecules). This method was repeated for tetramantane, however the choice of vectors was not obvious. Similarities between tetramantane and diamantane can be drawn at this point: both contain a C_3 rotation axis. Using this information, vectors were chosen to attempt to recreate the vectors used for the diamantane analysis, as can be seen in 3.21.

An initial analysis of these vectors (figure 3.22) produced graphs that showed a range of behaviours. Most molecules exhibited signs of switching from one orientation to another, undergoing a 120 degree rotation along the C_3 axis. This was observed by seeing three distinct regions when the vectors over the entire simulation were plotted, and have been shown in figure 3.22. This was to be expected, however instead of obtaining only three states, 6 different states were found. This can be explained by dividing the 6 states into 2 sets of three. Both sets would contain a C3 rotation, and the two sets would be linked by a C2 rotation perpendicular to the C3 axis.

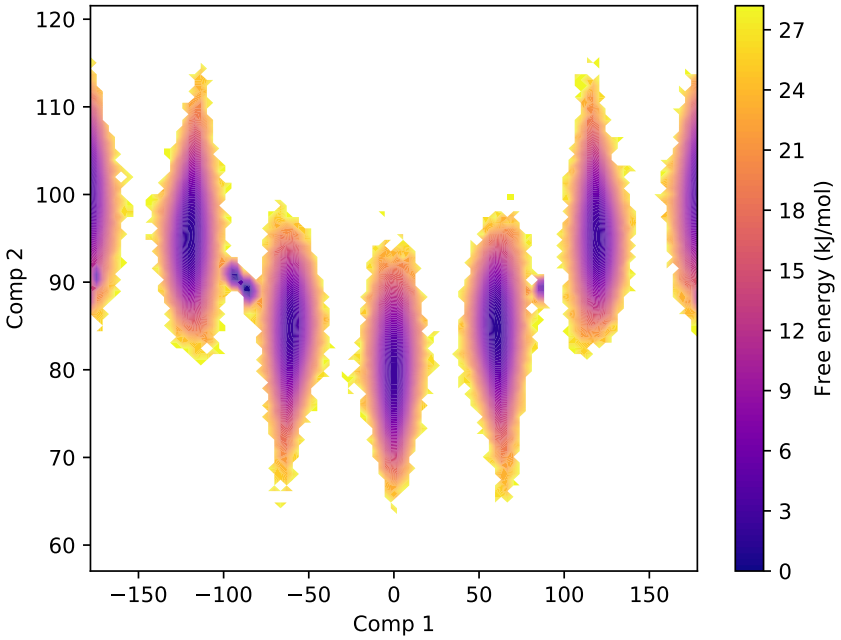

FIGURE 3.22: The free energy diagram generated from using the raw tetramantane vectors. As can be seen, 6 major states are available. This plot was generated using data from the 310 K simulation, and components 1 and 2 refer to the ϕ and θ values for the first vector from figure 3.21 .

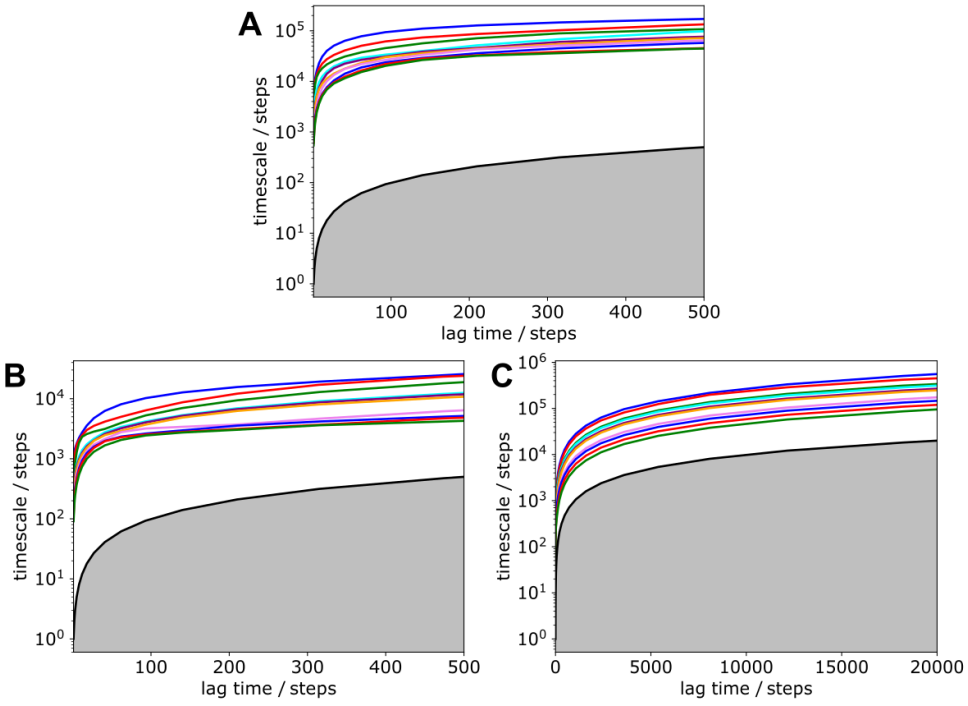

FIGURE 3.23: The implied timescale plots for tetramantane at A) 310 K B) 350 K or C) 400 K.

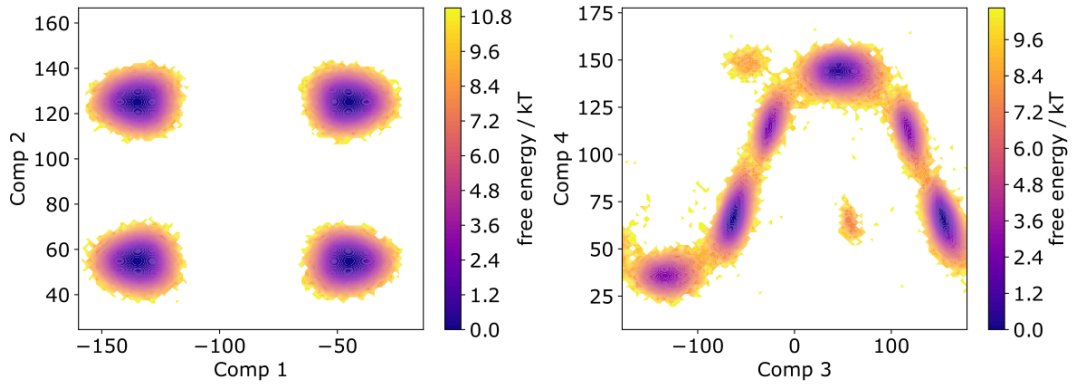

FIGURE 3.24: The output of directly plotting the vectors generated from the diamantane trajectory. The different component numbers refer to the specific vector components input, starting with vector 1's ϕ and θ angles and proceeding to vector 2, as shown in figure 3.12. This figure was produced using diamantane molecules with the same starting orientation.

All graphs show a rapid increase in the timescales when changing the steps within the first hundred, but the timescales level off after this period, instead of showing a steady rise. Expanding the graphs to further than 500 steps shows much the same behaviour, and so 300 steps were used to generate the MSM for all three temperatures of tetramantane.

3.4 Results

3.4.1 Diamantane

Analysing the trajectories took place in several stages. Firstly, the vectors of each molecule (Figure 3.12) over the simulation was plotted, to ascertain what could be seen from that. An example of these figures is shown in figure 3.24.

From this, three distinct minima can be seen. These can be attributed to the three orientations around the C3 axis that diamantane can rest in. Performing the same analysis for the overall system produces four larger areas corresponding to each orientation of the molecule, with each area containing three

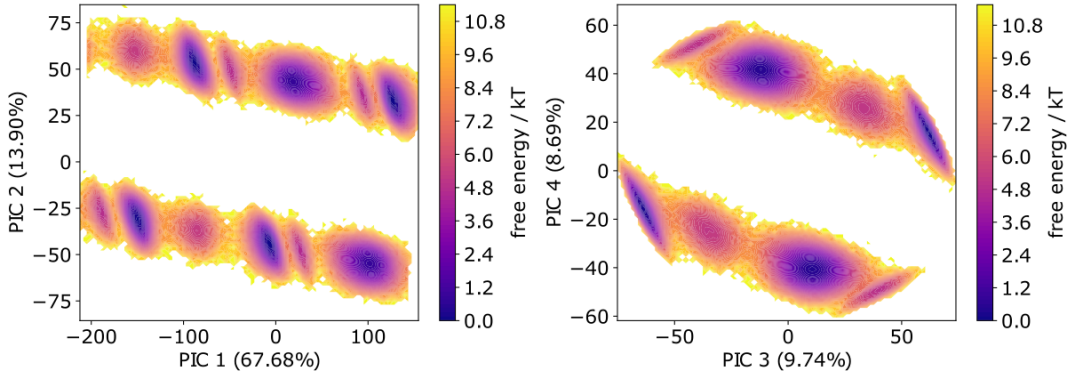

FIGURE 3.25: The four principal components for diamantane at 300 K. This data was generated by analysing all four positions together.

different minima. These initial figures support the data gathered from the NMR, but extracting more accurate details of each of these supposed states requires further analysis. The vector lists were subsequently input into the analysis code, and analysed in four ways: each orientation separately, the whole system, and then the same again but using the combined data from every temperature. During the PCA analysis, the data is transformed from "real" space (a physical description of system) into eigenspace, which is an energy surface derived from a function of the physical points. Moving between points in the eigenspace is similar to moving in real space, although moving along one axis in the eigenspace may result in moving along multiple axes in the real space. As the analysis produces a different eigenspace for each set of data, analysing the combined temperature data, and then extracting the trajectories relating to each individual temperature allows direct comparisons between the temperatures to be made.

300 K Simulations

Figure 3.25 shows the four components of the diamantane PCA for every molecule. While plotting the PCA components for a single orientation is

State from or State to	0	1	2
0	-	7.01±0.24	6.87±0.25
1	7.36±0.26	-	7.04±0.23
2	7.36±0.27	7.17±0.22	-

TABLE 3.1: The mean first passage times from state to state in ns for the 300 K system.

clearer (as seen in figure 3.27), plotting all orientations together gives us insight into the motion between, rather than within, the orientations. As can be seen, twelve distinct states are available, each represented by a block of probability density. This is most clearly shown in figure 3.25. These correspond to the four initial arrangements of the diamantane, and the C_3 rotations within each of them. These are arranged into two bands, one with a positive value for PIC 2 and the other with a negative value. Considering the "facing" of the four diamantane molecules in the unit cell, these bands make physical sense, as two molecules are pointing in one direction, with the other pair pointing oppositely. However, the components created are not physical: There is no direct link from a component to a particular arrangement of atoms, it is simply a descriptor.

Twelve blocks of probability density, while expected, is not wholly useful for this analysis. Understanding the transition times between within each state is required to obtain a clear and useful picture of the full dynamics of the system. As a result, the system was split into 4 different positions depending on their starting arrangement: A, B, C and D. Each group was analysed individually, undergoing each step in the analysis separately. Taking position A as an example, the components generated from the PCA analysis were plotted:

As can be seen, there is a trail of probability density across both pics, with three overall minima and three local minima available. This trail can thus be divided into three states, each state representing one C3 orientation, and the positions clustered according to which state they are closest too. The

smaller, local minima would therefore refer to the intermediate states during the rotation. While we can see three minima are present in the free energy diagrams (figure 3.27), determining the number of metastable states to divide the Markov model up into requires a look at the ratio of the implied relaxation timescales. Plotting the ratio of the xth timescale to the $x+1$th timescale can show us how many states to use when creating the HMM. This has been discussed in sections 2.5.3 and 2.5.4. The resulting plot is figure 3.26. Performing this analysis allows us to construct Markov models for the transitions from state to state, which gives us an average time of transition of 7.1 ± 0.3 ns, with an energy barrier of 17.4 ± 0.5 kJ mol^{-1}. These uncertainties were estimated from the range of transition times.

Similar diagrams are produced for each of the other positions in the unit cell, and are shown in figure 3.27. Including the other positions in the analysis gives us the same energy barrier within uncertainty of 17.83 ± 0.8 kJ mol^{-1}. The analysis has already proven the hypothesis from the NMR studies: The presence of three distinct minima in the diamantane rotation, with a transition time and energy barrier that closely follows that predicted from the NMR.

Higher Temperature Simulations

At 310 K, we are still seeing only three minima for each starting orientation. However, the regions between them are becoming increasingly well-sampled, as the molecules now have the energy required to spend more of the simulation time in this regions. Accordingly, the transition time between states decreases to approximately 5 ns.

Continuing to raise the temperature to 325 K has the effect of allowing quicker transitions between the states, lowering the transition times to around 3 ns.

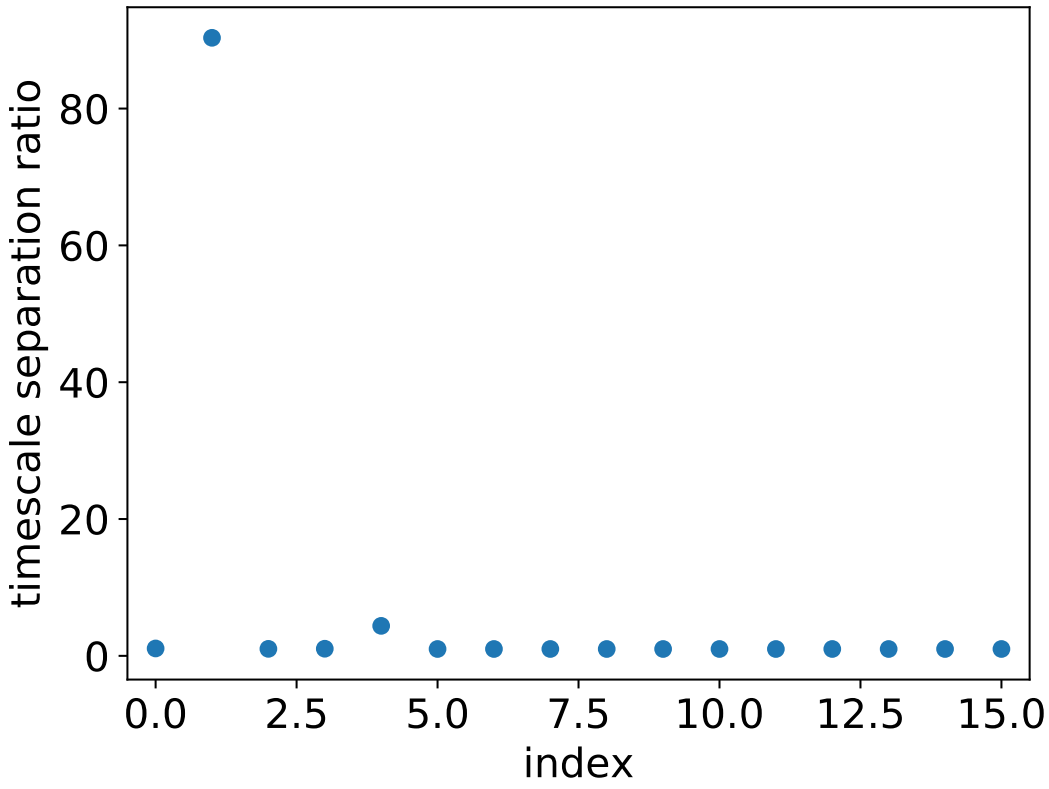

FIGURE 3.26: The ratios of the xth tmescale to the $x+1$th timescale. From this figure we can see that there is a large separation at index 1, which is the ratio between the 2nd and 3rd timescales, as the index numbers start at 0. This confirms that three metastable states is appropriate for the hidden Markov model construction.

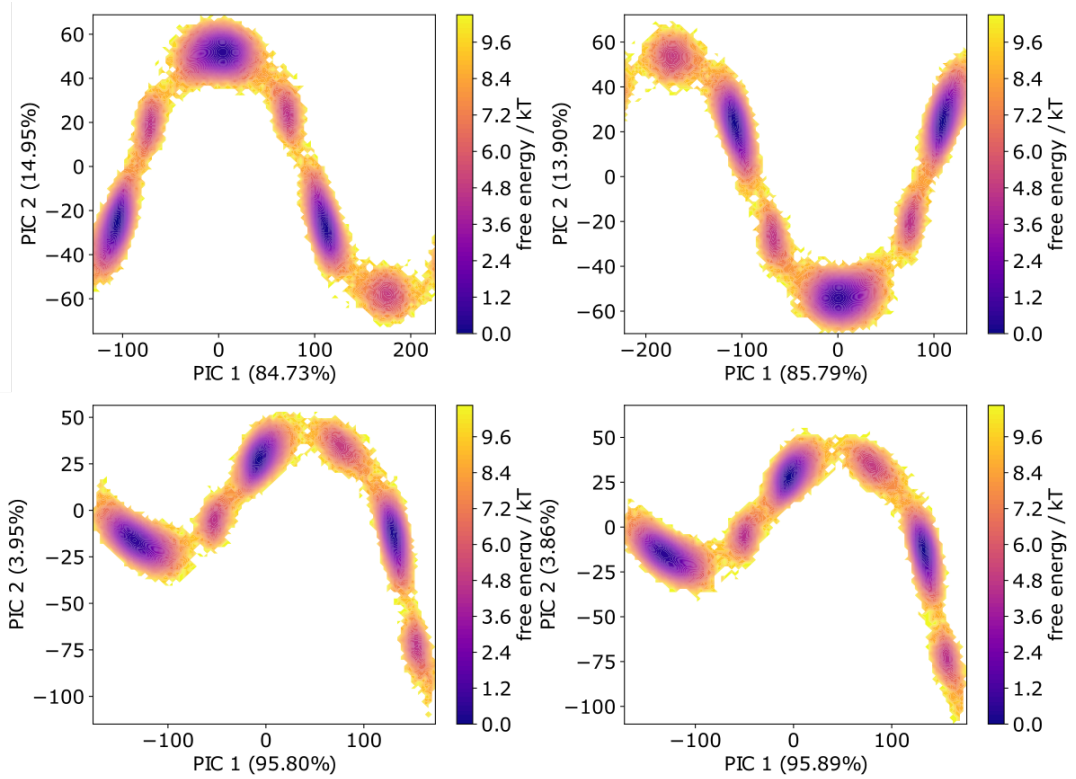

FIGURE 3.27: The first two principal components for each position at 300 K.

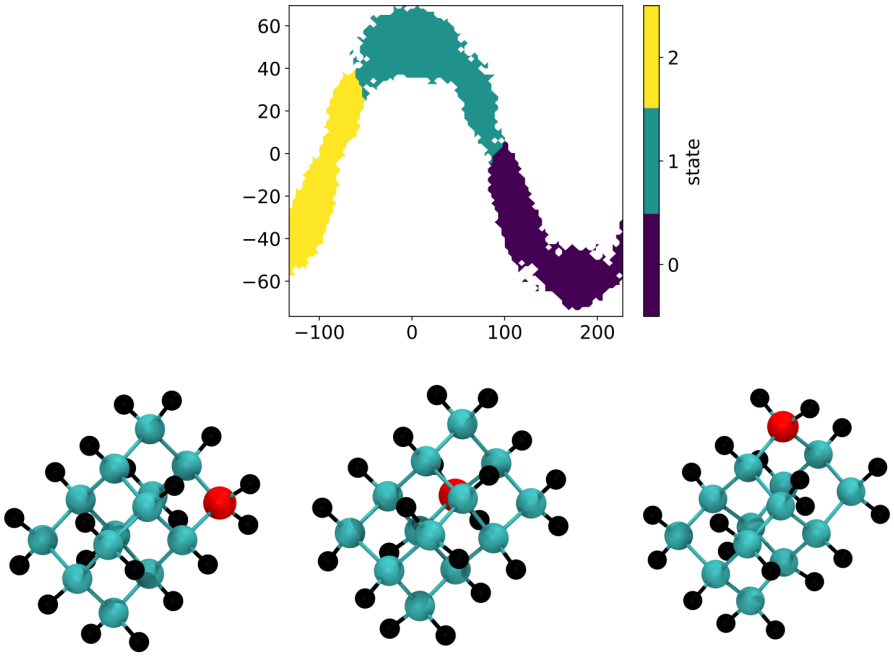

FIGURE 3.28: The state map for position A of diamantane at 300 K. The FED has been split into regions representing each of the three states, with the corresponding orientations of the diamantane presented below.

As we increase the temperature to 350 K, the transitions between states has dropped to around 1.3 ns.

To relate the temperatures together, all sets of data were run through the same PCA analysis at the same time, producing PICs that describe the two systems together. In this way, both sets of data can be plotted on the same eigenspace, allowing meaningful conclusions about the similarity of the data to be drawn. Figure 3.32 shows the data at 300 K and 350 K for one starting orientation plotted on the same eigenspace.

In addition to determining the rate of transition between the various states, we can also determine whether this motion is locally correlated, that is, whether a flip from one molecule triggers a flip in nearby molecules. To do this, the correlation coefficient was calculated using the clustered data, with any two residues having a correlation coefficient of greater than 0.5 being recorded.

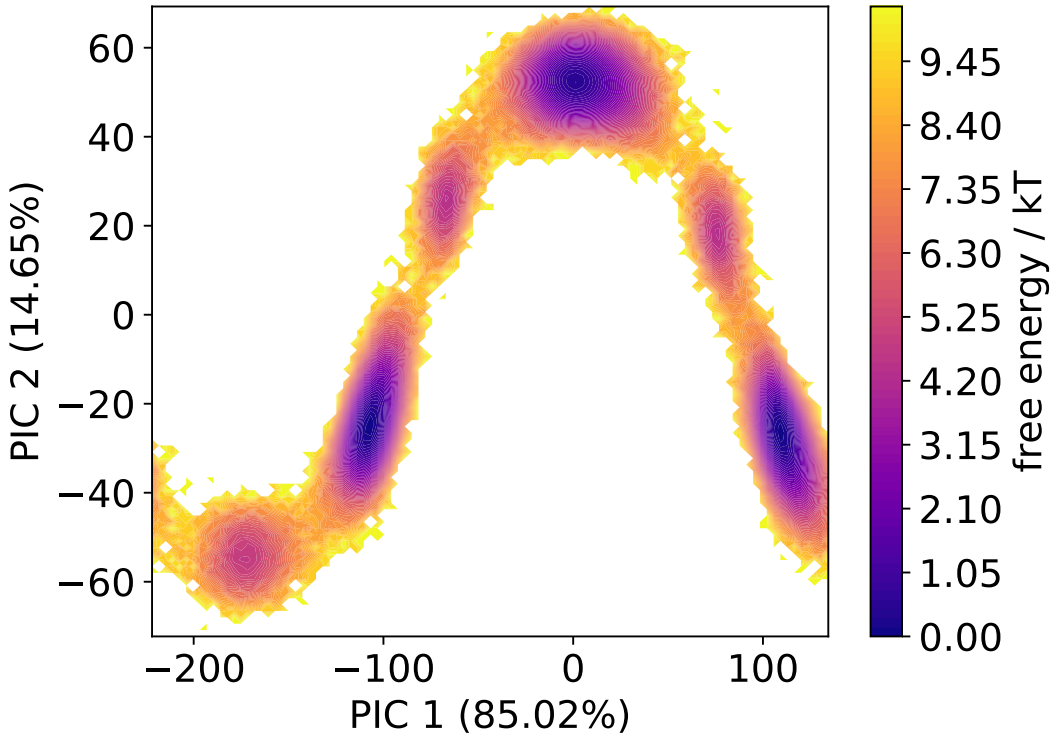

FIGURE 3.29: The first two principal components for position A of diamantane at 310 K.

Generally, correlation coefficients of over 0.8 indicate a strong correlation, so this analysis would pick up weak correlations between molecules as well. The resulting data proved showed that the molecules were almost entirely uncorrelated, having coefficient values of smaller than 0.1. This is a good indicator that the processes involved in the rotation are entirely random.

3.4.2 Triamantane

From the TICs plots (figure 3.33), we can see that as the temperature rises, the number of accessible states increases. At low temperatures, each molecule has lower energy, and is unlikely to rotate, resulting in two highly populated states, one for each starting position, and two lower populated states. At higher temperatures, the molecules are more free to rotate, resulting in four accessible states, each with similar levels of population (see figure 3.35). The

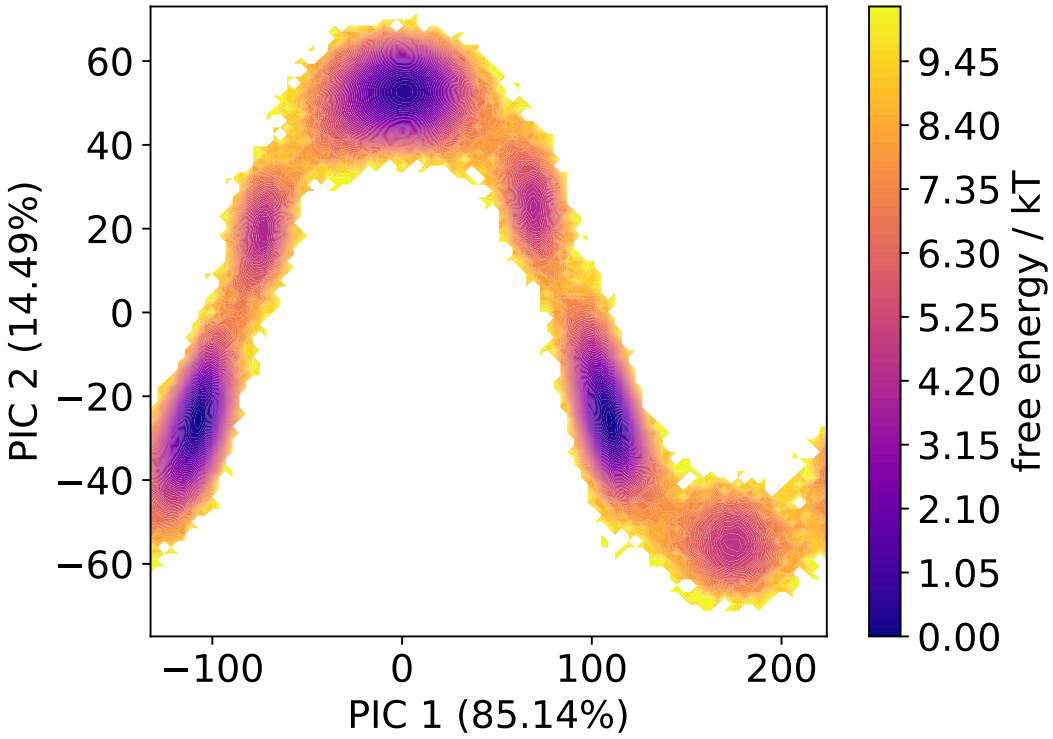

FIGURE 3.30: The first two principal components for position A of diamantane at 325 K.

rotations are very slow regardless of temperature, resulting in an incomplete model at lower temperatures, whereas higher temperatures provide enough data for a Markov model to be constructed. By dividing the system into two starting states, we can show differences between the rotations. Lower temperatures show some limited rotation between two different states, and extracting these shows us that rotation around the C_2 axis is possible, albeit very slow. As the temperature increases, we see the four states become apparent, indicating motion around our theorised pseudo-C_2 axis is taking place, although the actual form of the motion may not match our theory. Evidence for the transitions between the states is shown in figure 3.35.

The three sets of TIC plots (figure 3.33) show differing behaviours for each temperature. At the lowest temperature, we see two areas of probability density, indicating that even at these low temperatures, transitions between these

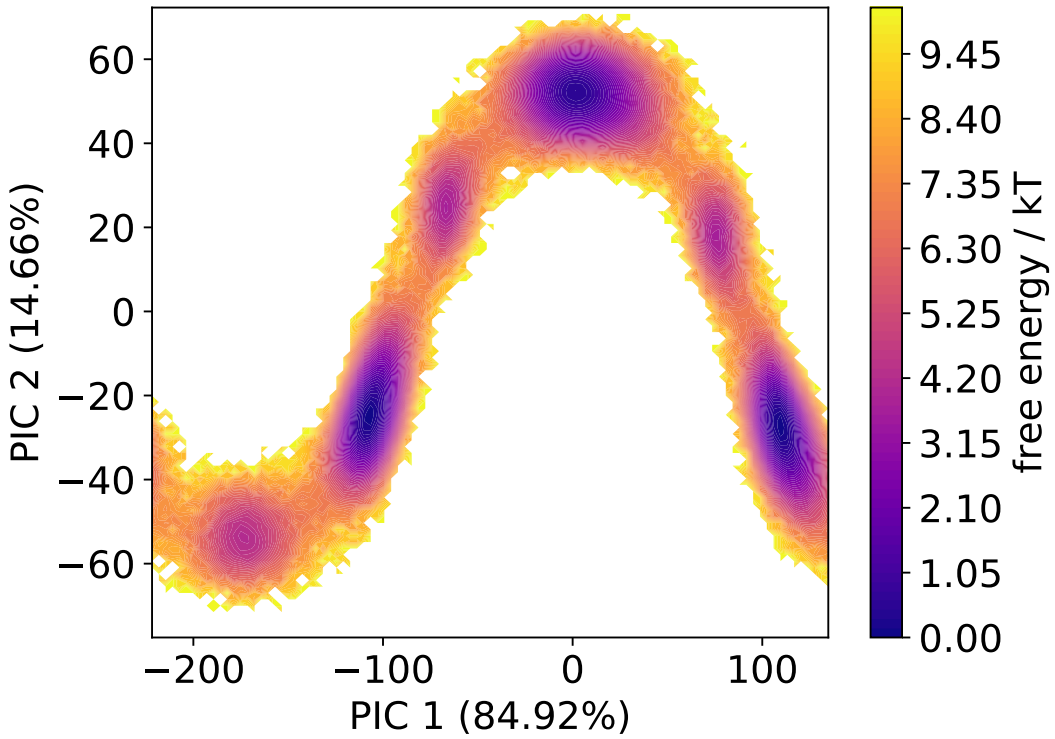

FIGURE 3.31: The first two principal components for position B of Diamantane at 350 K.

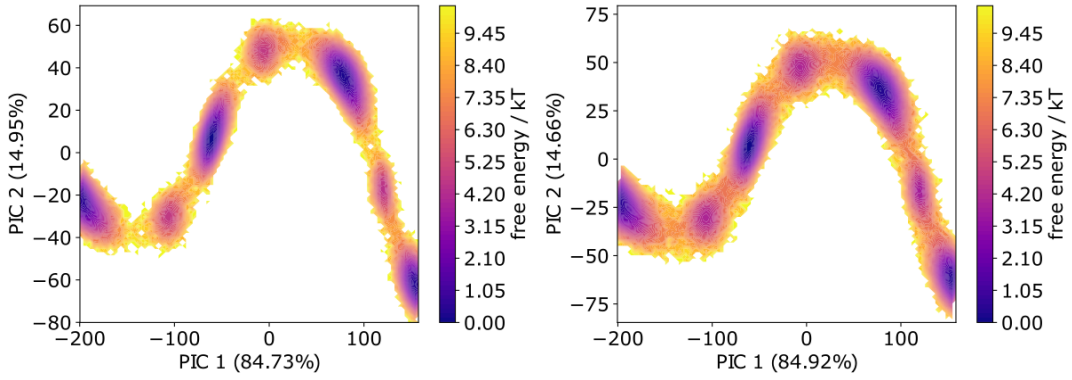

FIGURE 3.32: The first two principal components for diamantane at 300 (left) and 350 (right) K. This is the result of putting all temperature data through the same PCA analysis, to ensure they are all on the same eigenspace.

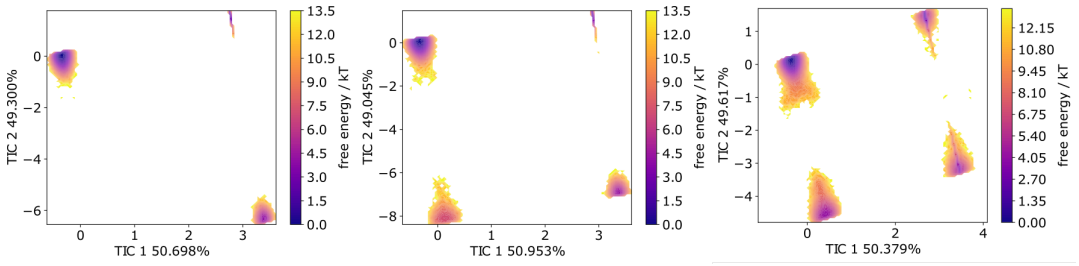

FIGURE 3.33: The TIC 1/2 plot for triamantane for position A. From left to right, the plots are for 310 K, 350 K and 400 K.

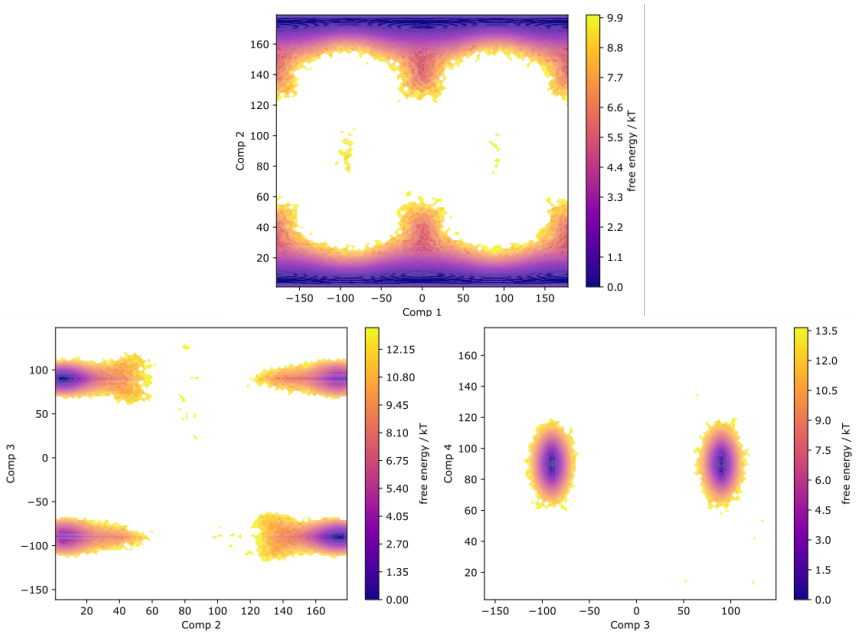

FIGURE 3.34: The free energy diagrams for the raw vector data for triamantane at 400 K. Each component refers to the four ϕ and θ vectors obtained from the molecules, with components 1 and 2 referring to the first vector in figure 3.17 and components 3 and 4 referring to the second vector.

states are possible. As we increase the temperature, the two areas of density fill out, indicating that the molecule is now able to explore a greater area of the probability space. However, while producing an MSM for this temperature was possible, coarse-graining it into an HMM proved difficult. As can be seen from the ITS plot (figure 3.20) for 350 K, the timescales of motion are still within the grey area, meaning the motions described by those timescales are unable to be observed through the use of an HMM. Performing a Boltzmann inversion of the data produces figure 3.34. The figure shows four states, as hypothesised, with unexplored areas between them. To determine the transition times between states, we can continue our analysis and construct an HMM. Only the 400 K simulations provide enough data on the rotations for a reasonable HMM to be produced, with four states being chosen. The four states are shown in figure 3.35.

The states clearly show the motions available to the triamantane within the crystal, both the C_2 and the pseudo-C_2 rotations exhibited by moving from state 2 to state 3. Transitions between the states are fairly infrequent, with transition times of 3 microseconds as average for the C_2 rotation and transition times of 24.5 microseconds for the pseudo-C_2 rotation. This data is obtained from transition path theory analysis of the generated HMM, while the transition time matrix for 400 has been shown in table A.5. Additionally, the energy barriers between such states are also high, with barriers on the order of 21 kJ mol^{-1} for the C_2 rotation and 31 kJ mol^{-1} for the pseudo-C_2 rotation.

3.4.3 Tetramantane

For tetramantane, we started by analysing the data using TICs. The resulting plot for TIC 1/2 has been presented in figure 3.36, and the plots for the other temperatures have a similar appearance. From these plots, we can see that

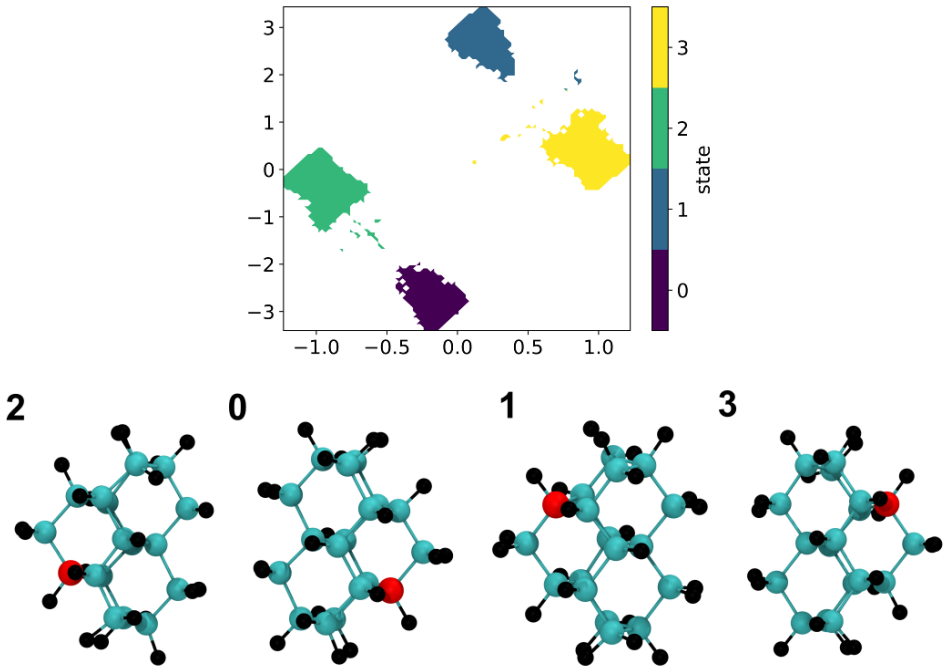

FIGURE 3.35: The state map and associated state samples for triamantane at 400. This figure shows the four regions on the tics plot that refer to each state, with an example of the corresponding orientation below each state. From left to right, the orientations correspond to states 2, 0, 1 and 3. From this, we can see that to transition between state 2 and 0, we need to rotate around the pseudo-C_2 axis, whereas to transition between state 2 and 1, we rotate around the C_2 axis. Transitions between 2 and 3, or 1 and 0 could indicate the presence of another pseudo-C_2 axis however.

State from\State to	0	1	2	3
0	-	16±5	1.2±0.7	9±3
1	-	-	-	-
2	-	15±5	-	7±3
3	-	8±3	-	-

TABLE 3.2: The transition times from state to state in μs for the 400 K system. Any transition time that does not exceed 150% of the uncertainty has been omitted. The full table is available in the appendices. The high uncertainties on many of the values is indicative of the slow speed of these transitions, and the lack of transition events present in the simulation. The simulation lasted 400 ns, and yet the model is predicting transition times of at least one order of magnitude above this.

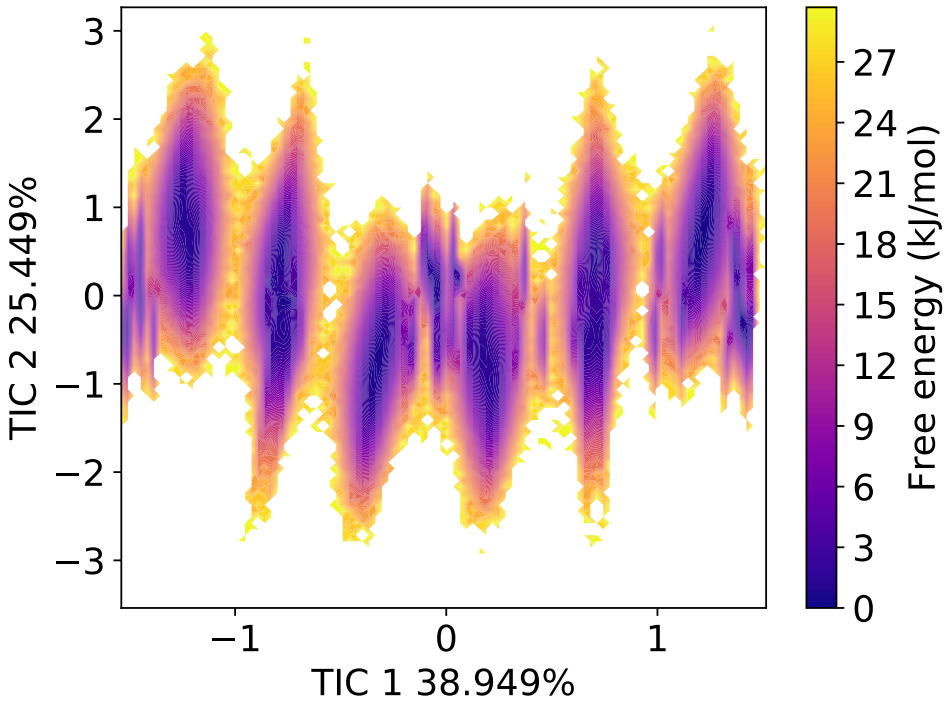

FIGURE 3.36: The TIC 1/2 plot for 350 K for tetramantane.

there exist many areas of probability density, comprised of six large bands and smaller areas of low energy scattered around. From initial inspection of the molecule, we can estimate that 6 major states should be accessible, comprised of rotation around a C_3 axis and around one or more pseudo-C_2 axis. The TICs data shows six major bands, which could relate to these 6 theorised states.

When moving forward and attempting to generate the HMM for this system, it was found that separation into 7 states showed the highest ratio between timescales. The HMM was generated and state samples produced, however samples of orientations from each state showed no clear differences between them. The 6 states we expected to see from our theorised rotations were present within each state, in that extracting a handful of state samples for one state produced orientations that matched our theorised 6 states. As TICs have been primarily designed to follow slow motions, this could be an

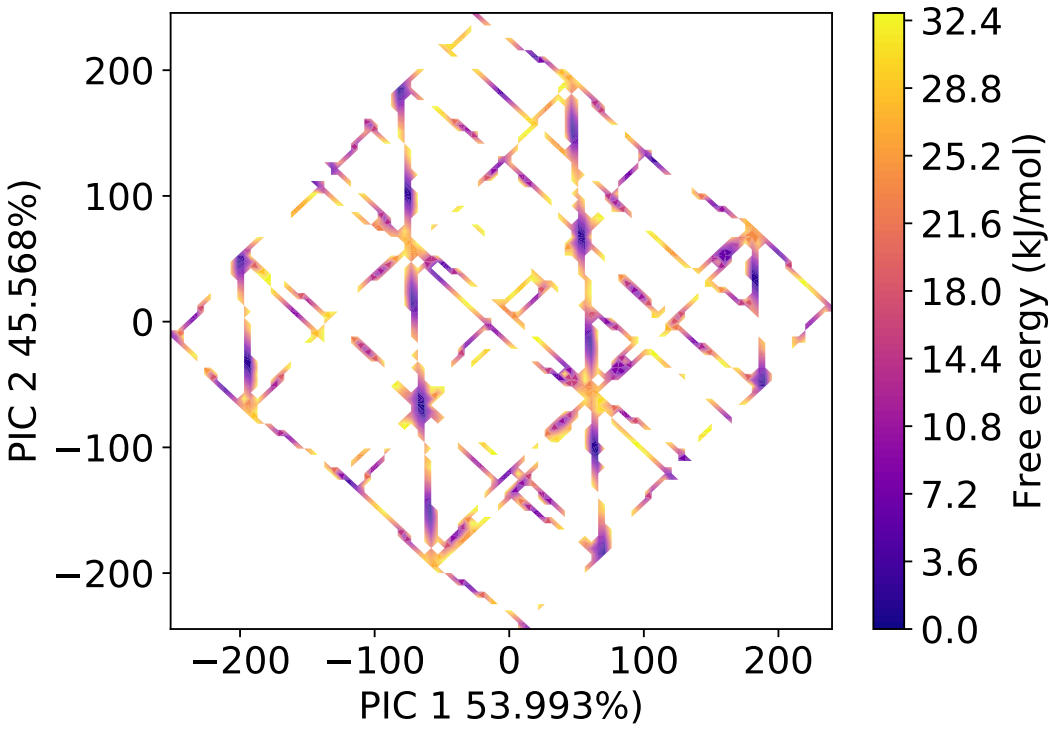

FIGURE 3.37: The PIC 1/2 plot for 350 K for tetramantane.

indicator that transitions between these states are rapid, and that a different method of analysis needs to be used. PCA is an excellent technique for studying rapid motions, and so further analysis was conducted using PICs rather than TICs.

The PICs plot above is rather less intuitive than previously generated graphs. Instead of a series of distinct states, it shows a web of potential states. Initial ideas based on the molecule's shape would suggest that a C_3 axis rotation is present, and can be mixed with a pseudo-C_2 rotation, similar to triamantane. The two highest PICs have a contribution to the overall motion of around 54% for the first component and around 46% for the second component. The third and fourth components have negligible contribution to the overall state space, with less than half a percent each. As the relationship between the

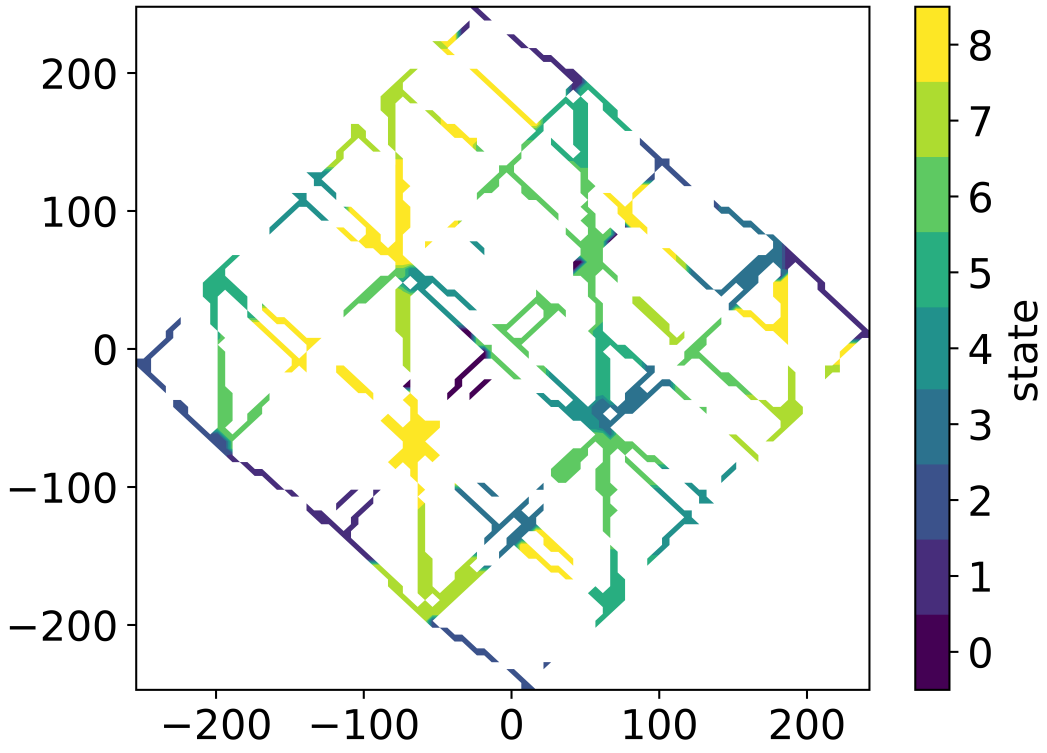

FIGURE 3.38: A statemap for tetramantane at 350 K. This was generated using 9 states, as suggested by the timescale separation plot. The scale here is discrete, with each state's region represented by a different colour on the map.

motion and the PICs is not physical, the resulting graph does not necessarily need to be intuitive, and so clustering and MSM/HMM generation proceeded. After generating the MSM and coarse-graining it into an HMM, the following statemap was produced.

Rather than the 6 states thought to be available, the largest seperation between groups came by grouping into 9 different states. The statemap above shows the splitting of the PICs into the various states, while the diagram below shows the labelled state examples for each. Each state appears to be spread throughout the pic-space, indicating that a range of disparate pic values are grouped together into one state. Extracting examples of each of the states yields something interesting, as shown below.

As can be seen (figure 3.39), the 6 states initially theorised are present in the

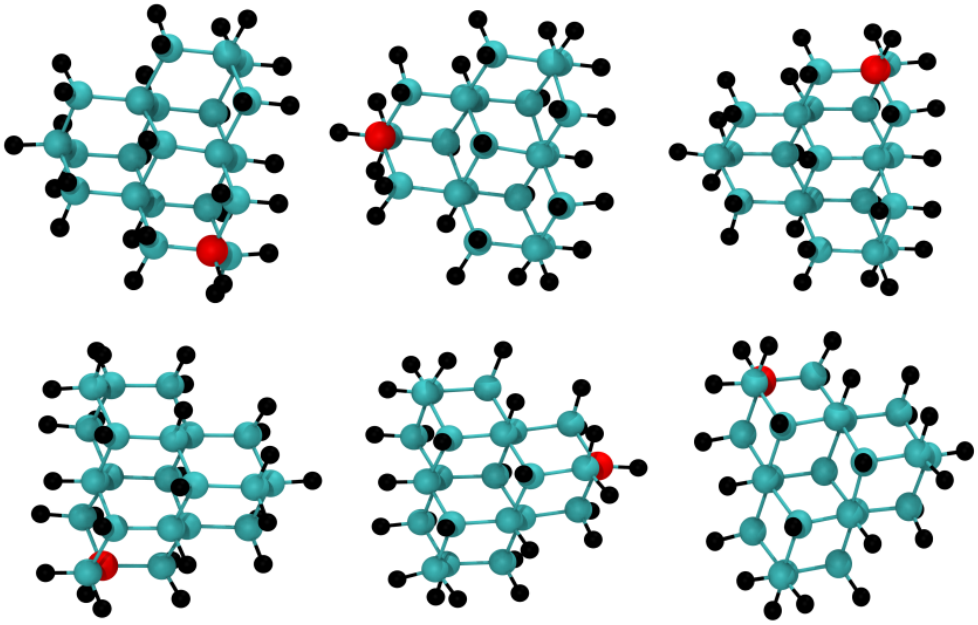

FIGURE 3.39: Example states for 350 K for tetramantane. While the analysis resulted in dividing the trajectory into 9 different states, visual inspection of the sample orientations generated by the code resulted in only 6 visually distinct orientations being found, with other states showing only minor adjustments of the overall orientation.

example states, however all of them have been grouped into a single state. This behaviour is mimicked for each of the states, meaning that the analysis, while it has clearly shown the presence of the states theorised, is picking up on some other form of motion in the data. By comparing the pic plot and state map, we can see that barriers between states vary widely, with a maximum barrier of around 30 kJ mol^{-1} present as the transition between some states.

Performing the correlation analysis on the data can give us a different insight into the motion. Correlated motions have previously been studied through the use of metadynamics [44], as accelerated MD methods have been used as alternatives to MSM methods [141]. In the case for tetramantane, this analysis is not performed by splitting the data into states, but uses the direct output of the PCA analysis, and shows us how correlated the various molecules are in their motion. Extracting the correlation coefficients for each temperature shows us that as the temperature increases, there are fewer examples of pairs of molecules achieving a coefficient of higher than 0.5. At 310 K, there are roughly 250 different pairs of molecules that have a coefficient higher than 0.5, with this number dropping to around 75 at 350 K, while at 400 K there are no longer different pairs of molecules with this behaviour. This could show that motion within the system at low temperatures is concerted, that is, once one molecule has gained enough energy to move, its motion allows neighbouring molecules to copy it. At higher temperatures, the motion becomes less linked, as each molecule can overcome the energy barriers associated with motion on their own.

3.5 Conclusion

In this chapter, we analysed dynamics in solid systems, using the diamantane system as our first model and progressing onto triamantane and tetramantane. Using previous NMR knowledge to guide us, a method has been

developed to probe and observe solid state dynamics using MD techniques, and the results have been shown to match well with the experimental data, replicating activation energies and reproducing the behaviour inferred from the NMR experiments, specifically the C_3 rotation. The diamantane results showed simple three-state motion at the lower temperatures, and more complex dynamics at the higher temperature showing that our method is able to detect the different motions of the system using the same input. Identifying features of the molecules or system that can be used to describe it is a key step, and the use of intramolecular vectors has provided the results in this case.

This method can be easily adapted to other systems, as the initial input is simply a list of any property over time, and can be altered to fit any reasonable description of the system. Further validation of the method is required however, to test its applicability to a range of systems. To achieve this, the same method was attempted on the triamantane and tetramantane systems.

The analysis of triamantane and tetramantane has shown a few of the limitations involved with these methods. Equilibration of the starting systems has been shown to be key to ensuring accurate data. The choice of PCA or TICA analysis has been explored, depending on the timescales of the observed motions. The triamantane system was shown to skip between four different states, related together by a C_2 and a pseudo-C_2 rotation, with energy barriers of around 21 and 31 kJ mol^{-1} respectively. Transition times between these states varied widely, however an average rate of 12 μs was obtained for the C_2 rotation, and 4.6 μs for the pseudo-C_2 rotation. This data was only obtained during the high temperature simulations, as lower temperature runs did not allow the system to travel freely between the states. The timescales of motion observed by the MD simulations fit in nicely with the timescales observed through the NMR experiments.

The tetramantane system proved limited in the data we could obtain. While the analysis was successfully run on the system, useful data only arose at high temperatures as well, suggesting motion is fairly prohibited. The correlation coefficient data for 350K suggested that the motion is highly correlated, with various molecules following each others progress through their rotations.

Based on the data obtained from the simulations, the NMR data can be interpreted further. As the simulations show two motions within the diamantane system, further NMR experiments can investigate this. In particular, while the energy barriers calculated from the C_3 rotation closely follow those from the NMR, calculating an energy barrier and transition rates for the 90° rotation would determine whether this can be seen via NMR, as well as providing a new pathway to allow the molecule to relax. Additional relaxation measurements should keep these two distinct motions in mind. For triamantane, molecular motion at the microsecond level has been shown by the simulations, giving credence to the claim that the line broadening is due to motion at the rate of ^{1}H decoupling. Additionally, seeing the four states during the simulations would mean that any rate measurements from the NMR would need to keep in mind that the rate seen could be a combination of both motions available to the molecule. The molecular dynamics is providing a physically plausible model here that cannot be derived directly from just x-ray diffraction or NMR, but is explaining both finding there. Utilising this model is key to assessing which experimental data is sensitive to the motion, as well as showing both the MD and NMR are in quantitative agreement. For the tetramantane, the simulations see six distinct orientations, and the NMR provides evidence of a C_3 rotation. Further analysis would need to consider the possibility of the molecule flipping through a mirror plane, although this MSM analysis is struggling with this system. Improving the

analysis or tailoring it further to this system will provide more information that can be used to influence NMR experiments. Altering the temperature of experiments could freeze this motion out, and is an avenue that could be explored in future NMR experiments, although the relaxation data may be insensitive to the flips observed.

4 Furosemide-Picolinamide

4.1 Introduction

Furosemide (FS) is a loop diuretic drug, commonly used to treat fluid build-up due to heart, liver or kidney failure, but is also used to treat high blood pressure [142]. It works by inhibiting the NKCC2 transport protein, binding the chloride transport channel, causing a loss of sodium, chloride and potassium ions in urine to incite the body to remove fluids. The solid form of FS has low aqueous solubility and low biological membrane permeability, meaning the bioavailability of the drug is limited [143, 144]. Due to strong intra- and intermolcular hydrogen bonding, the crystal resists dissolution, and so modifying these properties can enhance the solubility of the drug and improve its bioavailability and effectiveness.

One method to increase the bioavailability is to design coformers and solvates that can increase solubility and drug release rate. FS is a flexible molecule with a furan ring, and contains both hydrogen bond donor and acceptor groups, with additional halogen bonding available from chlorine. All this combined allows FS to exist in a number of different solid forms when crystalline. In this chapter, I will focus on the study of one particular FS co-crystal, a furosemide-picolinamide (FSPA) co-crystal with solvates formed from either ethanol (FSPA-ETH) or acetone (FSPA-ACE). FSPA forms a co-crystal containing solvent channels, the structure of which has been shown in figure reffig:fspascheme.

The structure from X-Ray Diffraction (XRD) suggested the solvent molecules present were highly disordered. The structure of the solvated co-crystal was determined to contain a channel between the molecules, in which the solvent lies. While electron density was observed in the solvent sites, resolving this into the atoms of the solvent proved fruitless, as the solvent was too disordered to be modelled. The methodology presented in the methods chapter seems perfect for identifying the dynamics and extracting relevant properties, with SS-NMR allowing us to verify these results.

4.2 NMR Results

Previous NMR experiments were recorded by Hannah Kerr of the Hodgkinson group at Durham with the intent of characterising the disorder of the co-crystal. Focusing on the ethanol solvate, an attempt to use the ^{13}C linewidth of the ethanol CH_2 peak was made to provide information on the dynamics present. While an increase in linewidth was observed, the data was not good enough to extract an E_a from the data. Instead, the ^{13}C T_1 relaxation times of the ethanol peaks was used, measuring the decrease in the relaxation times to fit to an Arrhenius curve. This yielded an E_a value of 19 ± 0.9 kJ mol^{-1} for the CH_3. The fitting has been shown in figure 4.2.

Studies on the FSPA-Acetone system included lineshape analysis at a variety of temperatures, the result of which has been shown in figure 4.3. The fitting between the -31 °C data and a 500 kHz flip rate is reasonable, however the other temperatures have proven difficult to assess, as the lineshapes could apply to a large range of flip rates. The use of ^2H T_1 experiments to predict an activation energy yielded a result of 7.9 ± 0.2 kJ mol^{-1}, and the fitting has been shown in figure 4.4. These data points give us a starting point to work with, as well as a set of experimental data to compare to.

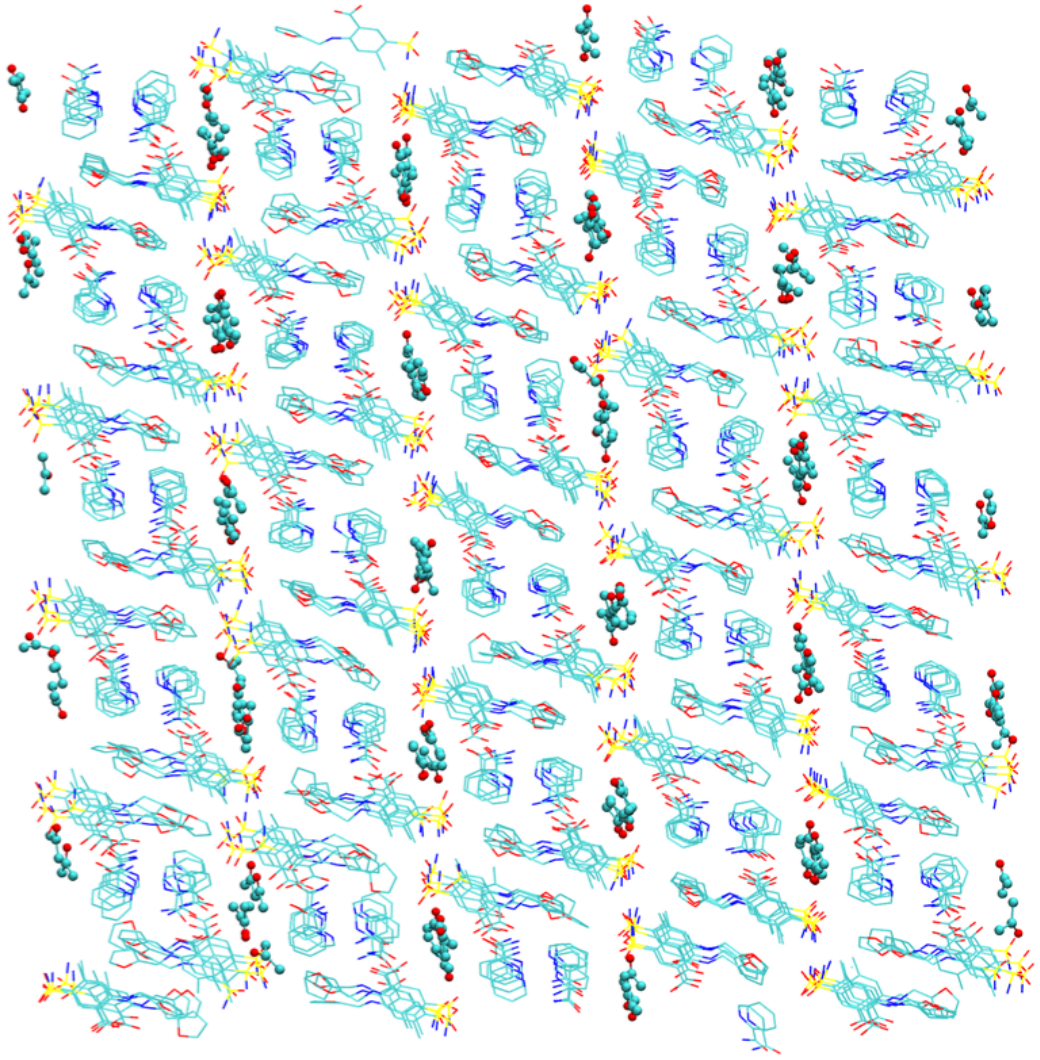

FIGURE 4.1: A top-down view of the FSPA crystal simulation. The furosemide and picolinamide molecules have been represented by lines, while the solvent (acetone) is clearly shown sitting in the solvent channels, which continue through the plane of the paper.

4.3 Particulars of Method

This study considers the FSPA solvate crystal structures containing ethanol or acetone. The system was built using coordinates from supplied CIF files obtained by XRD at 120 K. The unit cell of FSPA-ACT at 120 K was 5.047 Å x 14.428 Å x 14.778 Å with 77.35, 83.13, 96.34 degree angles, and FSPA-ETH was 5.209 Å x 14.641 Å x 14.565 Å with 76.38, 87.59, 83.24 degree angles.

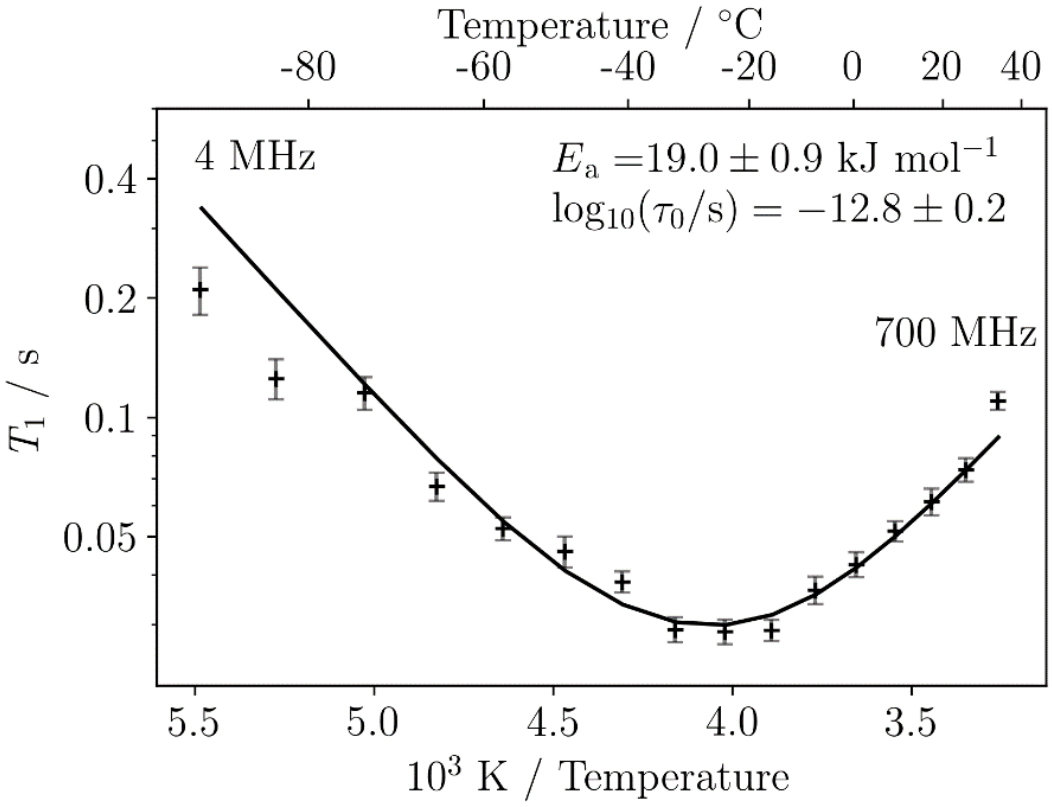

FIGURE 4.2: ^{13}C T_1 relaxation times of the ethanol solvent fit to an Arrhenius model. Data supplied by Helen Wickins [145].

Atomic positions of solvents were not resolved through XRD, therefore we placed one molecule of solvent (either ethanol or acetone) in the channel of each unit cell of FSPA channel.

All of the systems have been modelled with Amber force field [147, 148], assigned through GAFF [149, 150]. Molecular dynamics simulations were performed with GROMACS 2016.4 [111]. Each simulated system was first energy minimised using a steepest descents algorithm, with convergence when the maximum force on any atom was less than 100 kJ mol^{-1} nm^{-1}. Then, to allow solvent molecules to move into equilibrium positions, the FS or PA molecules were positionally restrained with the SHAKE algorithm [151, 152], while the solvent molecules were allowed to move freely. The system was allowed to equilibrate for 50 ns in the *NPT* ensemble with the velocity-rescale

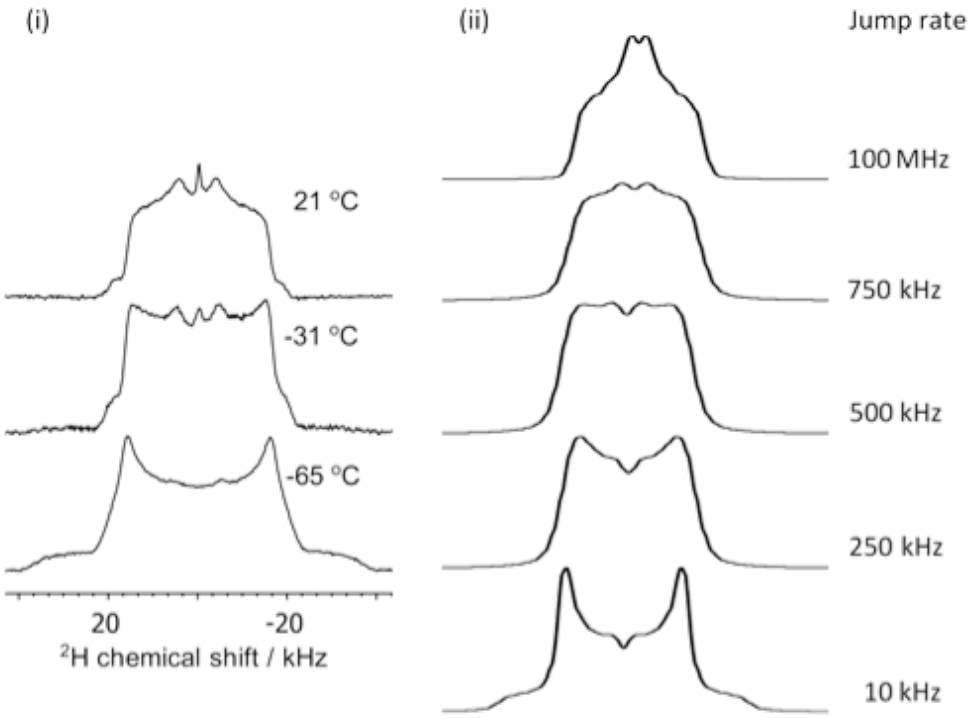

FIGURE 4.3: Variable temperature ^{2}H spectra of FSPA-Acetone compared to EXPRESS simulated static ^{2}H spectra using a two-site jump model. Data supplied by Hannah Kerr [146].

Berendsen thermostat at 120 K, the temperature coupling constant set to 0.1 ps, and an isotropic Berendsen barostat at 1 bar, with a pressure coupling constant of 1 ps. The solvent molecules were assumed to be relaxed when the solvent molecules were evenly distributed with respect to channel symmetry. This was determined by utilising the vectors shown in figures 4.5 and 4.6. The vectors can point either "up" or "down" the channel, and once the populations of "up" and "down" were equal, the system can be said to be equilibrated. This was judged qualitatively by eye by plotting the distribution of the relevant vectors. The convergence of the system was also assessed using RMSD. The RMSD of the system with respect to itself was plotted over time, and once this value had levelled off, the system can be said to be equilibrated. This is because as the simulation starts, the system will move a little

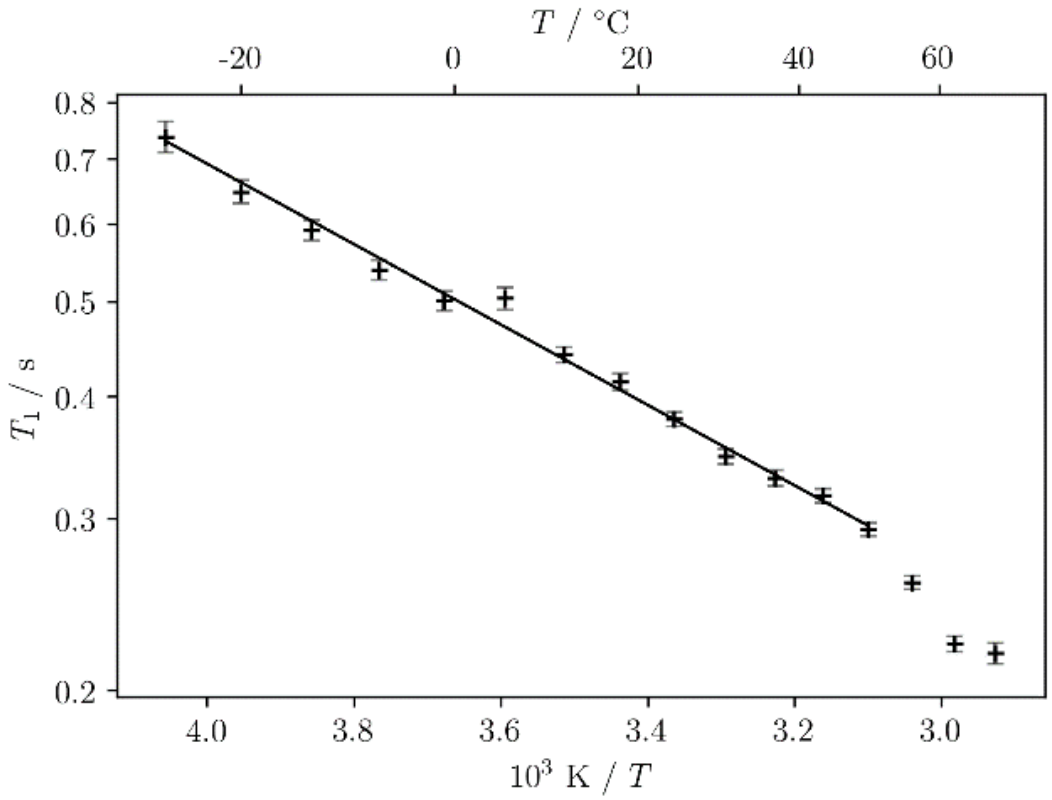

FIGURE 4.4: ^{2}H T_1 relaxation times of the acetone solvent fit to an Arrhenius model. Data supplied by Helen Wickins [145].

from the crystallographically fixed positions, but should maintain the overall structure. Once this relaxation has tailed off, the RMSD will level off, and the system is equilibrated. The positional restraints were then removed, and the system was allowed to equilibrate for 50 ns with the same protocol as above, with temperature set to 120 K as in the original crystal structure determination. The equilibration protocol described above allowed crystallographically poorly resolved groups to undergo reorientation and explore the available phase-space. The equilibration simulations were followed by an annealing run, using the same parameters as above, where the system was heated from 120 K to 323 K, with incremental temperature rises of 3 K/ns up to 273 K, and 1 K/ns thereafter, each step followed by 10 ns relaxation to prevent a lag, before the next temperature increment. Using this simulated annealing, we extracted structures at temperatures of 250 K, 263 K 273 K, 283

K, 294 K, 303 K, 313 K and 323 K. These structures were then simulated in an *NPT* ensemble with the Nose-Hoover thermostat at the given temperatures with the temperature coupling constant set to 1 ps, and an anisotropic Parrinello-Rahman barostat at 1 bar, with a pressure coupling constant of 1 ps for 100 ns. The choice of these temperatures is related to previous NMR work, with spectra having been taken at 273 K and 294 K.

As with the previous systems, there was a clear choice of two vectors to describe the overall orientation of the molecules. For acetone, we took one vector from the central carbon to the oxygen, and another between the two methyl group carbons. For ethanol, we took one vector from the methyl carbon to the methylene carbon, and then from the methylene carbon to the oxygen atom. Figures 4.5 and 4.6 illustrate the analysis vectors chosen. By choosing these vectors, we are excluding any methyl dynamics from the acetone; these have been well-documented and the dynamics within the group is very fast, so is of little interest in this system. The ethanol vectors do exclude the O-H dynamics from consideration. This is for similar reasons however, as the O-H bond is much more mobile than the C-O bond. While the O-H bond could form a hydrogen bond with the crystal atoms, the same is true of the C-O bond vector, and the C-O bond in conjunction with the C-C bond give us a much greater description of the overall molecular orientation.

Following the same method for optimisation as detailed in Chapter 3, first the TICA lag time is optimised, followed by the number of cluster centres, MSM lag time and the number of HMM states. This process will be shown first for the FSPA-ethanol system, then for the FSPA-acetone system.

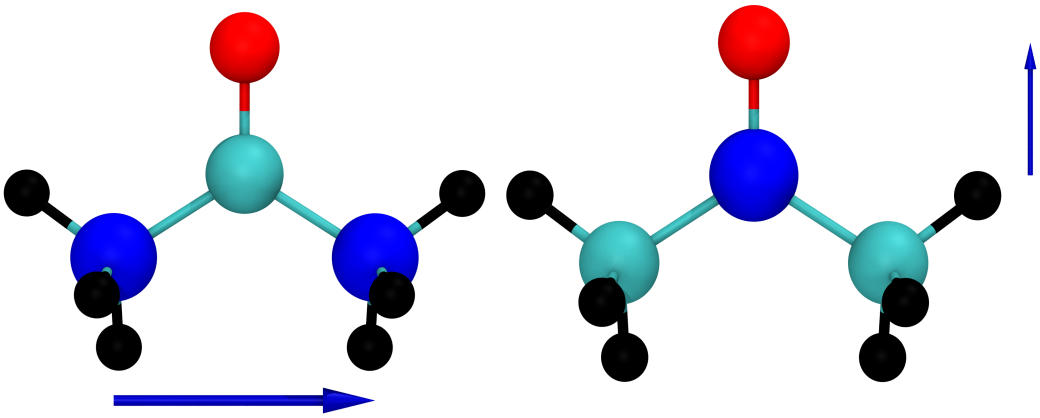

FIGURE 4.5: The analysis vectors chosen for the FSPA-Acetone system. The dark blue spheres indicate the carbons chosen for the vectors in each diagram, while the arrows show the direction and orientation of the vector.

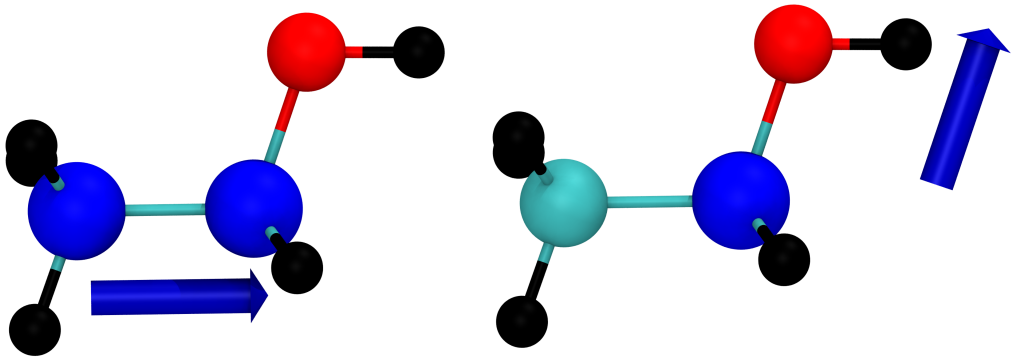

FIGURE 4.6: The analysis vectors chosen for the FSPA-Ethanol system. The dark blue spheres indicate the carbons chosen for the vectors in each diagram, while the arrows show the direction and orientation of the vector.

4.4 FSPA-ethanol

Graphs 4.7, 4.8, 4.13 and 4.14 show the implied timescales calculated after performing a TICA transformation using a variety of lag times (1, 10, 500) and clustering using 128 clusters. The lag times are simply multiples of the simulation timestep, which for these simulations is 0.5 fs. From the plots, it can be seen that the timescales continue to increase, but after only 3 steps start to become smaller than the lag time. Producing a Markov model while the

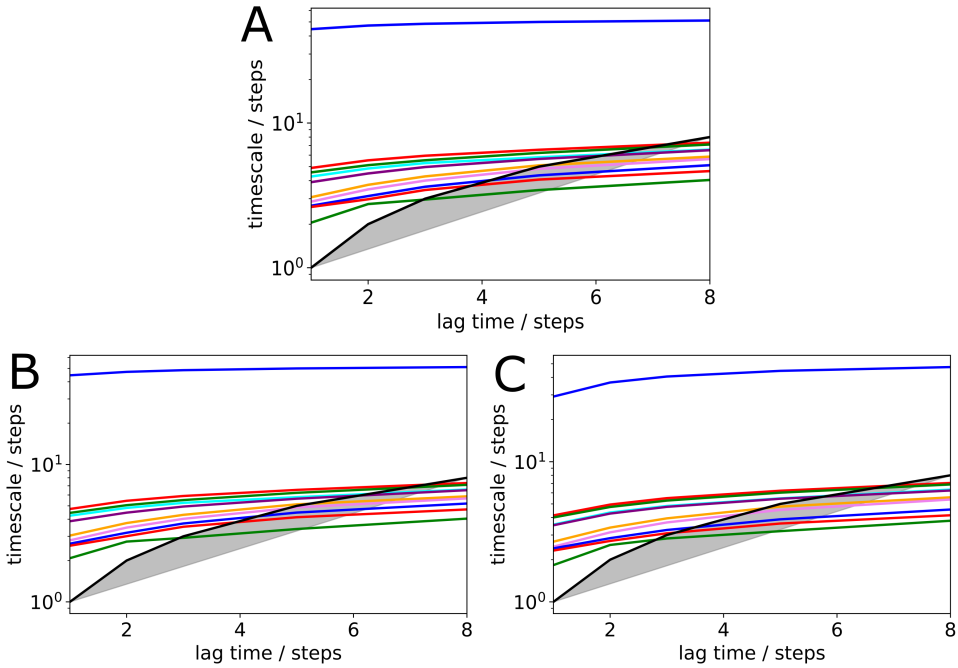

FIGURE 4.7: The implied timescale plots generated to test the effect of changing the lag time on the FSPA-ETH system at 273 K. The graphs were produced using lag times of A) 1 B) 10 and C) 500 simulation steps, where each step is 0.5 fs.

timescales of interest are smaller than the lag time chosen leads to inaccurate results. It can also be seen from the plot that the first implied timescale is far higher than any further ones, suggesting already that a 2-state model is best suited to fit this data. There is no major change in the timescales dependent on the TICA lag time, and so a lag of 5 was chosen.

The TICA data was then clustered with varying numbers of cluster centres: 12, 48, 64, 128, 256 and 512. Graphs 4.8 and 4.14 show the implied timescales for MSMs determined after using the specified number of cluster centres:

Increasing the number of cluster centres (see figure 4.8) does have a small effect on the timescale plots, as the lines become less curved, straightening out somewhat earlier on. This is because the transitions described by the lines become smoother, as rather than jumping rapidly between two poorly

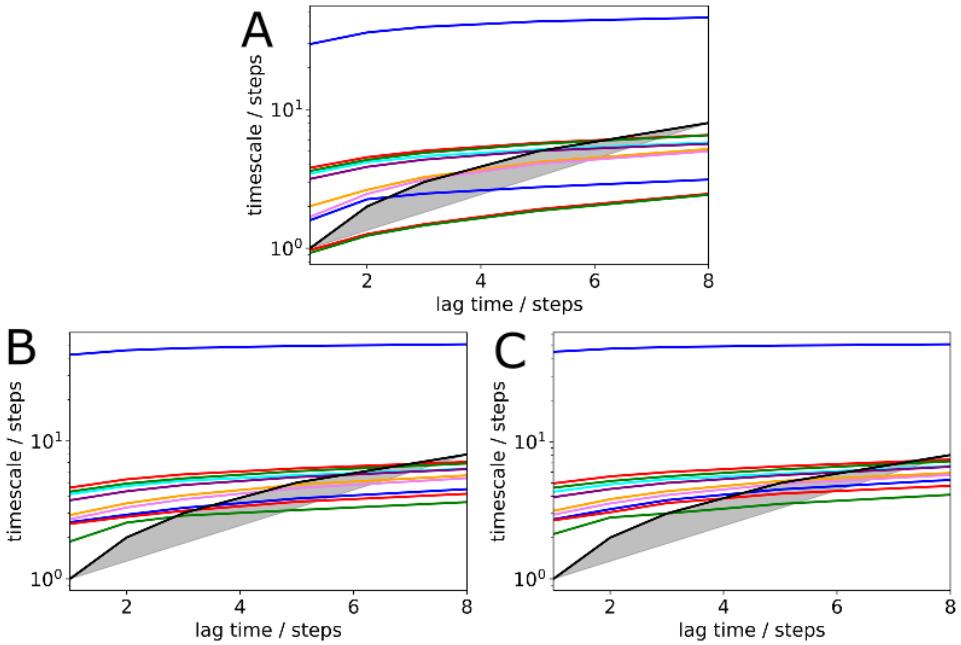

FIGURE 4.8: The implied timescale plots generated to test the effect of changing the number of cluster centres on the FSPA-ETH system at 273 K. The graphs were produced using A) 12 B) 64 and C) 512 cluster centres.

clustered states, the molecule's trajectory is able to explore more of the intermediate space between two states. This results in timescales converging earlier, and thus a smaller Markov time for the model. The Markov time is the smallest time at which the model converges, displayed on the graph by the beginning of the almost horizontal portion of the graphs. Choosing a smaller Markov time in this case is useful: the timescales soon become shorter than the lag time, meaning we start to lose useful information. By choosing a smaller Markov time we can include as much of the dynamic information as possible while ensuring the timescales are not shorter than the chosen lag time. These concepts have been discussed in sections 1.1.3 and 2.6.1.

Once appropriate parameters have been selected, we can start to analyse the data. The analysis is presented in detail in Chapter 3, and a similar process will be followed here. We first consider the simulation run at 273 K.

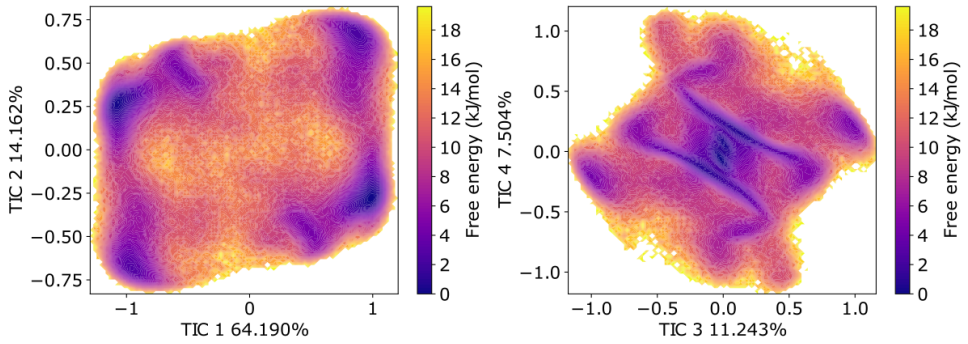

FIGURE 4.9: FSPA-Ethanol TICs 1 to 4. These graphs were produced for the 273 K system, using 256 cluster centres and a lag time of 10 simulation steps, where each step is 0.5 fs.

Figure 4.9 shows the free energy plots for the TICA components for the FSPA-ethanol system, the first figure showing the projection of the free energy on the first and second TIC, while the second figure does the same for the third and fourth TIC. From the graph showing components 1 and 2, we can see 6 major areas of low energy: two each on the bottom right and upper left, and one each on the upper right and bottom left. These six states can be split into two groups of three, with one group on the left, and the other on the right. Moving between states within the groups has an associated energy barrier of around 10 kJ/mol, while moving from group to group has a barrier of around 14-16 kJ/mol. These values are obtained from inspection of the free energy diagram (figures 4.9). We use the state map to divide the FED into regions, then move from the centre of one region to another, taking the path that requires the least amount of energy. In this way, we can plot a course from state to state and determine the rise in energy as we move between them. The presence of 6 groups of low energy also gives us an indication that choosing 6 metastable sets to produce the Hidden Markov Model (HMM) would be a sensible choice. The ITS plot can give us additional information at this point.

Figure 4.10 shows us that the timescales involved are extremely rapid at 273

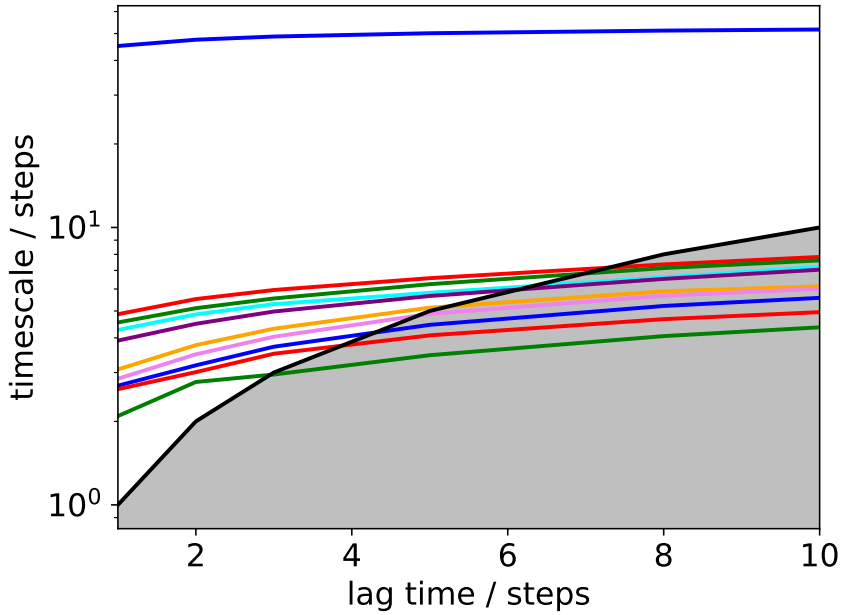

FIGURE 4.10: FSPA-ethanol ITS plot at 273 K. This graph shows how the length of each implied timescale changes as we vary the lag time of the Markov model. Each step is 0.5 fs. A TICA lag time of 10 steps and 256 cluster centres were used, matching graph B in figure 4.7 and placing it between graphs B and C in figure 4.8

K, with a lag time of around 3 reached before moving into the grey region, where the timescale is shorter than the lag time. Additionally, this plot shows us that the first timescale is far greater than any of the others. As the ratio of each timescale to the next one guides us in choosing the number of metastable states available, this indicates that partitioning the data into two groups would be best, which goes against our intuitive reading of the FED. However, this shows the usefulness of Markov modelling: this is telling us that one motion is much slower than the rest. We can confirm how many states would be better by producing a graph of the ratio of each timescale with its following timescale, presented in figure A.2 in the appendix. The timescale separation confirms what can be seen from the ITS plot: While there is a ratio of around 8 between the first timescale and the second, the ratios are just over 1 following that, showing that using 2 metastable sets for the HMM is recommended. This seperation into two states is obvious from the ITS plot, however performing this ratio analysis allows us to see if any further partitioning of states would be wise. After producing the HMM, we can create a metastable set graph, showing the density of each group on the TICA plots we have seen before. This graph is presented in figure 4.11.

Our initial guess is correct: The HMM has separated the trajectory into two groups, one on the left and the other on the right. These will correspond to two major orientations of the ethanol. Examples of these orientations can be extracted from the trajectory. These examples have been presented in figure 4.12.

The key difference between these states is the position of the alcohol group: in one state it is pointed up, while the other is pointed down, and the major transitions present in the system are between these up and down states. The transition time between states is rapid, with average transitions times of 0.5 ± 0.006 ns between states at 273 K. This number was obtained using

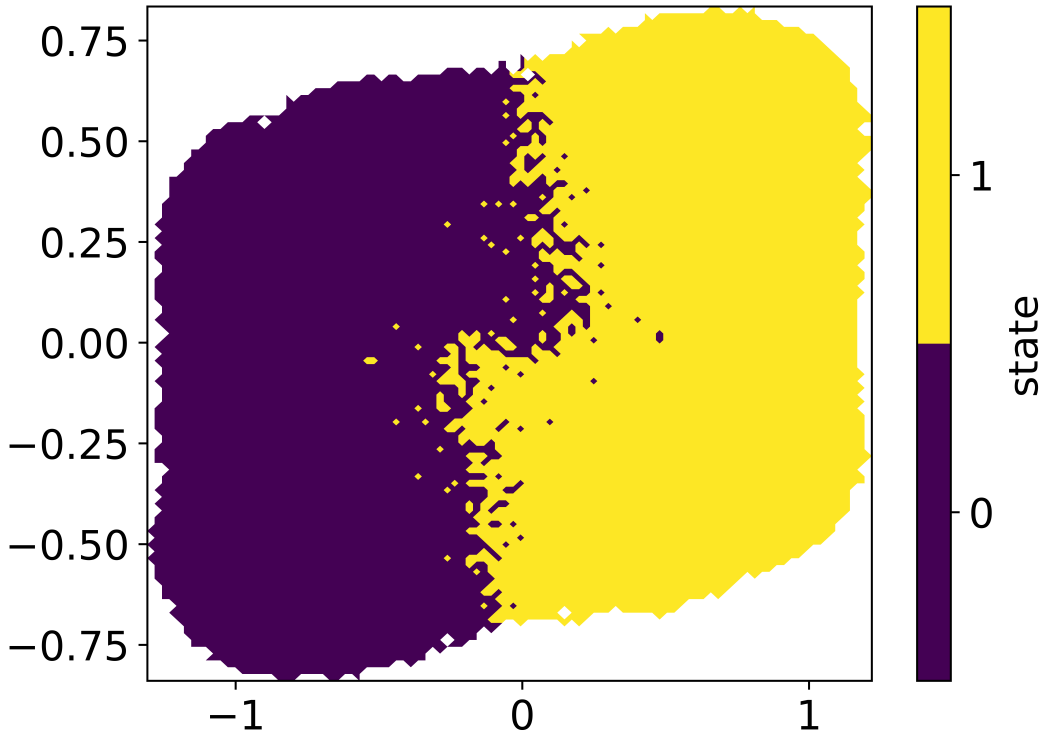

FIGURE 4.11: FSPA-ethanol state map using 2 metastable. This graph shows the division of the free energy diagram projected along TICs 1 and 2 into the different metastable states. The presence of points relating to one state within the area of another is a result of the clustering in the four dimensions given by the TICA analysis, but only presenting in two dimensions.

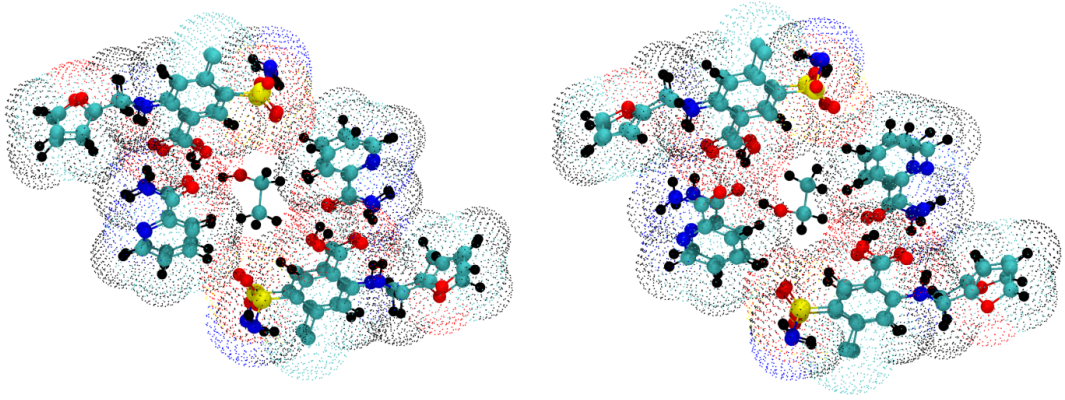

FIGURE 4.12: FSPA-ethanol example orientations.

State from\State to	0	1
0	-	0.535±0.006
1	0.585±0.006	-

TABLE 4.1: A table showing the times in ns of transitioning from any state to any other state for the FSPA-ETH system at 273K. Uncertainties are calculated based on the standard deviation of 100 estimates of the transition time. The lack of overlap between the times can be attributed to the trajectories exhibiting more transitions in one direction than the other. As this is just an artifact of sampling, extending the trajectory length would equilibrate these times.

Temperature	Average transition time
263	0.600±0.007
273	0.560±0.006
283	0.470±0.005
294	0.362±0.003
303	0.287±0.002
313	0.219±0.001
323	0.183±0.0006

TABLE 4.2: A table showing the average transition time between the two states for the FSPA-ETH system for each simulated temperature. These values were calculated by averaging the transition times from state 0 to state 1 and state 1 to state 0.

Transition Path Theory (TPT), as explained in Chapter 2. Analysis of the correlation coefficients show that no two distinct molecules shared a correlation coefficient of higher than 0.5. This indicates that the motion of the individual ethanol molecules is uncorrelated with the motion of the others.

Table 4.2 shows the average transition time between states for each temperature simulated. As the temperature increases, the transition time steadily decreases, starting to tail off. Modelling the evolution of these transition times as an Arrhenius type expression with the equation $\tau_c = \tau_0 \exp \frac{E_a}{RT}$ yields an activation energy of 14.2 ± 0.4 kJ mol^{-1}. This energy is roughly the same strength to a typical hydrogen bond, which could indicate the key process for the ethanol rotation is the making and breaking of these hydrogen bonds. From the NMR, an activation energy of 19 ± 0.9 kJ mol^{-1} was found, so the NMR and MD results are on the same order of magnitude.

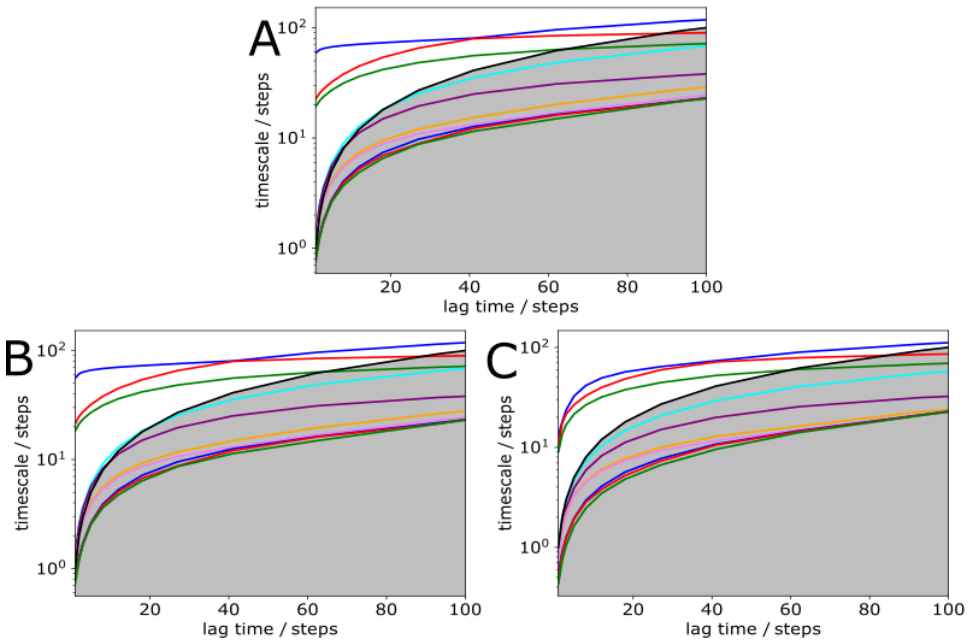

FIGURE 4.13: The implied timescale plots generated to test the effect of changing the lag time on the FSPA-ACE system. The graphs were produced using lag times of A) 1 B) 10 and C) 500 simulation steps, where each step takes 0.5 fs.

4.5 FSPA-acetone

The following graphs (figures 4.13 and 4.14) show the selected implied timescale plots calculated after performing a TICA transformation using a variety of lag times (1, 10, 500) and clustering using 128 clusters. From the plots, it can be seen that the timescales continue to increase, but after only 20 steps start to become smaller than the lag time.

The TICA data was then clustered with varying numbers of cluster centres: 12, 48, 64, 128, 256 and 512. Figure 4.14 shows the implied timescales for MSMs plotted after using the specified number of cluster centres: Increasing from 12 to 64 shows a significant improvement in the timescale, as the timescales move further away from the grey zone, with the highest timescale converging quicker than with 12. There is a similar improvement, albeit less dramatic, when changing from 64 to 256 clusters. No significant change was found when changing from 256 to 512, and so 256 clusters was chosen.

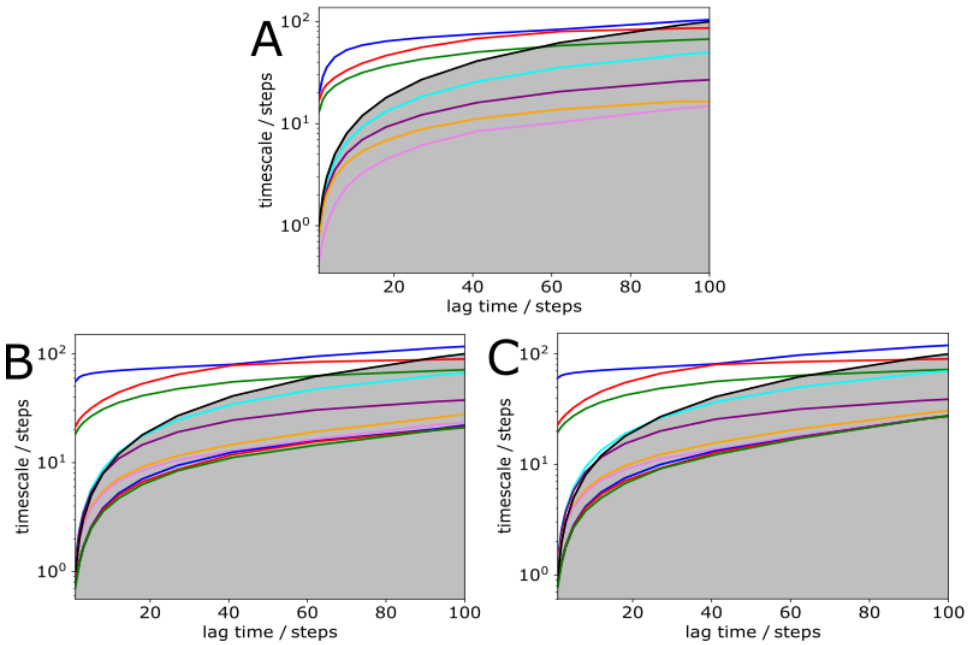

FIGURE 4.14: The implied timescale plots generated to test the effect of changing the number of cluster centres on the FSPA-ACE system. The graphs were produced using A) 12 B) 64 and C) 512 cluster centres.

From the NMR data, a simple 2 state flip was hypothesised for the acetone motion, as numerical simulations using this model were shown the match the experimental NMR well. Using this information to guide us, we can begin our analysis of the FSPA-Acetone system. First we will consider the simulation run at 273 K.

Figures 4.15 and 4.16 show the free energy plots for the first four TICA components for the FSPA-acetone system. From the graph showing components 1 and 2, we can see 4 major areas of low energy, one on each corner. This suggests that there are four metastable states available in this system. Moving between states has an associated energy barrier of around 12 kJ/mol in this case. This gives us an indication that choosing 4 metastable sets to produce the HMM would be an accurate choice. We can now turn to the ITS plot for additional insight.

The ITS plot shows us that the motion in this system is around 10 times

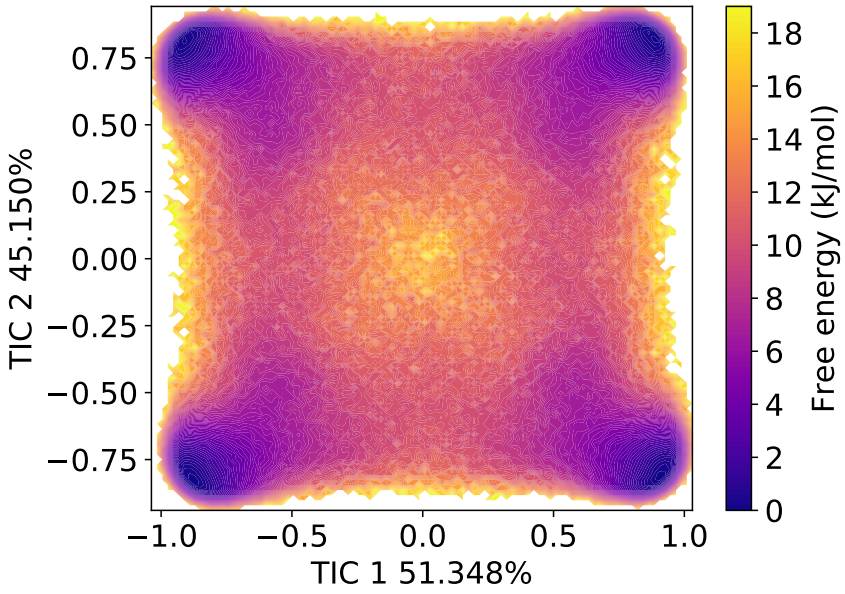

FIGURE 4.15: FSPA-acetone TICs 1 and 2. This graph were produced for the 273 K system, using 256 cluster centres and a lag time of 10 simulation steps, where each step is 0.5 fs.

slower than the majority of the motion in the ethanol system, although the slowest timescale in the ethanol system is comparable to the motion here. An appropriate lagtime from this graph is around 40 steps, compared to the 3 steps found for the ethanol solvate. Additionally, this plot shows us that the first three timescales are the easiest to access, with other, more rapid motions quickly falling into the grey region. From this, it confirms that 4 metastable sets would be appropriate for the HMM. We can confirm this by producing a graph of the ratio of each timescale with its following timescale, presented below.

The timescale separation differs from what can be seen in the ITS plot: While producing a HMM using 4 states (represented by index 2 on the graph) would be a good choice, using 6 states has an even greater separation ratio. However, the timescales in question lie in the grey region in the ITS plot, and so

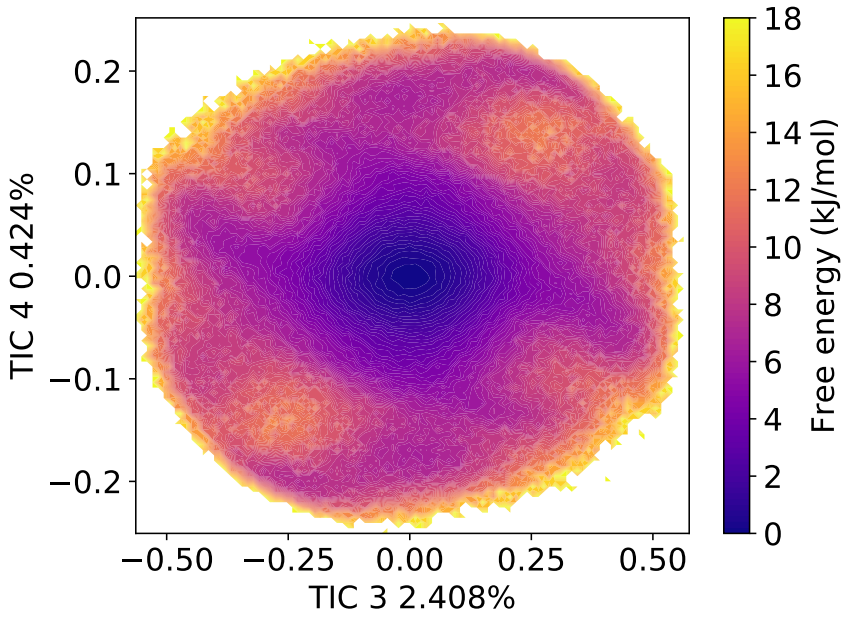

FIGURE 4.16: FSPA-acetone TICs 3 and 4. This graph were produced for the 273 K system, using 256 cluster centres and a lag time of 10 simulation steps, where each step is 0.5 fs.

using them for analysis must be performed with care. Producing two different HMMs, one for 4 states and the other for 6, would be prudent, to compare their features. After producing the HMM, we can create a metastable set graph, showing the density of each group on the TICA plots we have seen before. This graph has been presented in figures 4.19 and 4.20.

The state map plots show that while four states have been assigned to the four corners, as expected, there are two states that seem to be intermediates between pairs of points, one on the left of the graph and one on the right. These states are theorised to be the intermediate states between transitions. However, it is interesting that only two transitions have intermediate states, while the others do not. This could point to two options: either these transitions are slower, meaning the molecules spend more time in this stable intermediate state before moving to one of the four "main" states, or that these should also be included as "main" states.

From the statemap we can see that states 0 and 1 are the intermediate states.

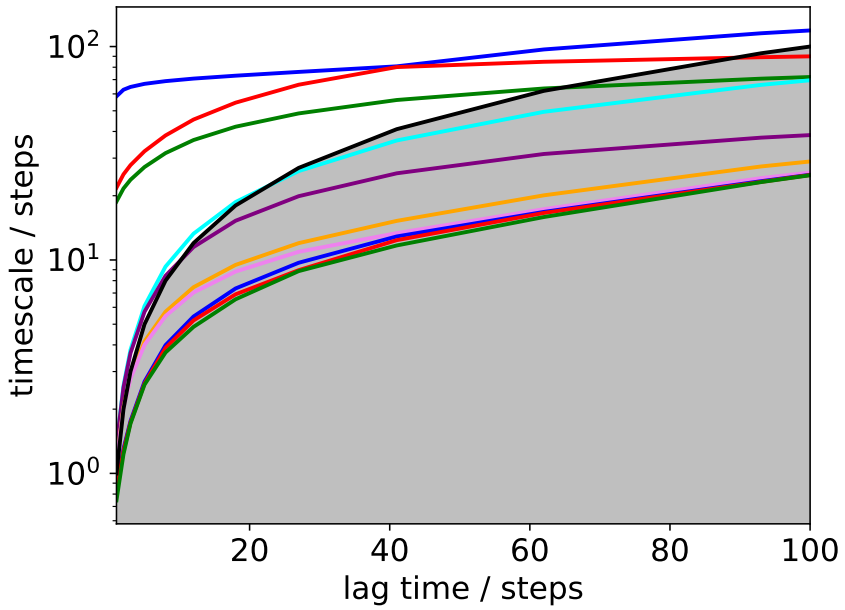

FIGURE 4.17: FSPA-acetone ITS plot. This graph shows how the length of each implied timescale changes as we vary the lag time of the Markov model. Each step is 0.5 fs.

State from\State to	0	1	2	3	4	5
0	-	2.283±0.176	4.792±0.213	4.618±0.246	4.306±0.255	4.374±0.240
1	2.197±0.190	-	4.597±0.211	4.462±0.241	4.487±0.255	4.580±0.255
2	3.759±0.324	3.652±0.284	-	4.668±0.289	4.626±0.215	3.474±0.281
3	3.743±0.321	3.674±0.282	4.826±0.274	-	3.301±0.312	4.726±0.193
4	3.573±0.328	3.841±0.278	3.761±0.183	3.442±0.283	-	4.541±0.294
5	3.559±0.340	3.851±0.269	3.691±0.272	4.784±0.206	4.459±0.327	-

TABLE 4.3: A table showing the rates in ns of transitioning from any state to any other state for the FSPA-acetone system at 273 K

Transitions between these states are faster than transitions out of these states by roughly 1 ns, while states 2-5 show fairly equal transition times between all states except those opposite each other on the state map (2-3 and 4-5). This suggests that the acetone moves from the "corner" states to an intermediate state, rapidly switching between the two intermediate states, and then settles into a more stable "corner" state again.

The state samples show us that the four states are related by two 180° rotations: One around the carbonyl bond and another perpendicular to this while still in the plane of the molecule. Combining these two 180° rotations gives

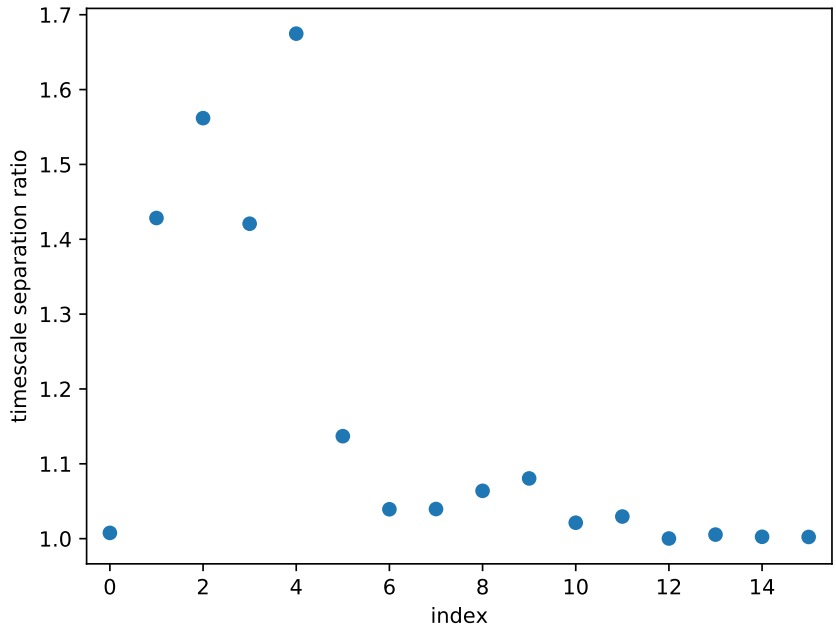

FIGURE 4.18: FSPA-acetone timescale separation plot. Each point shows the ratio of timescale n and timescale $n+1$, with index 0 showing the ratio between timescales 1 and 2.

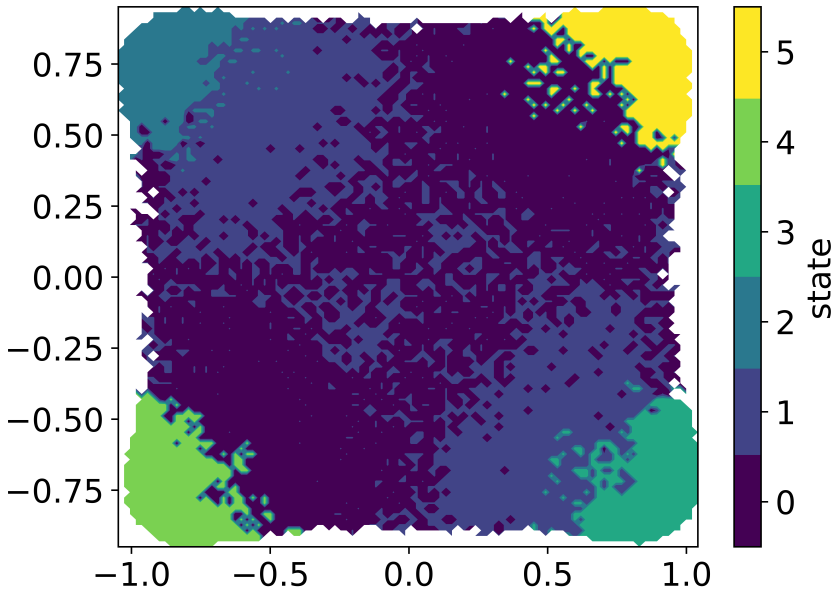

FIGURE 4.19: FSPA-acetone state map for 273 K. This graph shows the division of the free energy diagram projected along TICs 1 and 2 into the different metastable states.

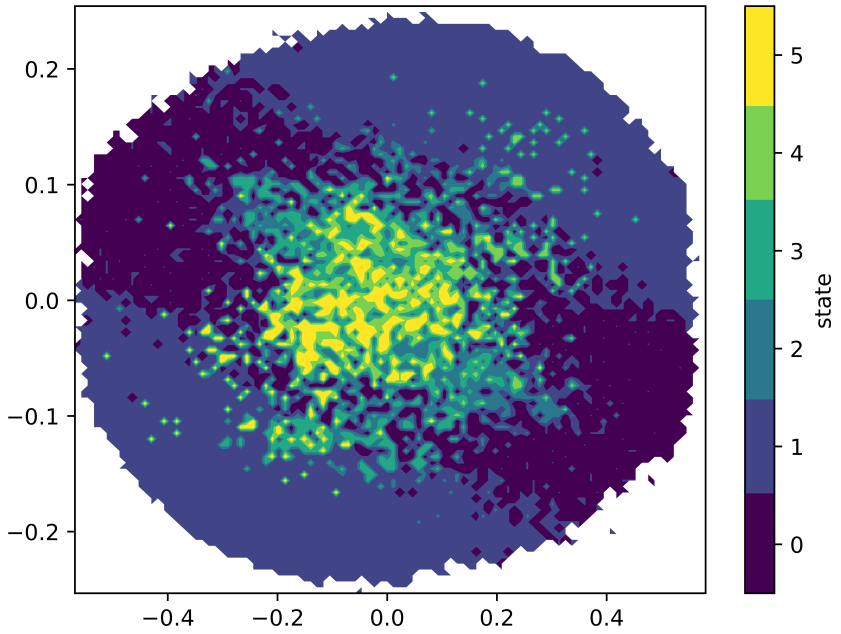

FIGURE 4.20: FSPA-acetone state map. This graph shows the division of the free energy diagram projected along TICs 3 and 4 into the different metastable states.

us the motions available to the molecule, with transitions times ranging from 4-5 ns at 273 K. These transitions times slow down significantly as you reach 250 K, with the time ranging from 15-22 ns. Correlating these states to the statemap shows us that transitioning from the "top" to the "bottom" corners involves a rotation about the C-C bond vector as shown in figure 4.5.

At temperatures higher than 273 K, the separation of timescales from the Markov model changes, indicating that division into 4 states is the optimum for the HMM. From the state map 4.22, we can see that the intermediate states from the lower temperatures have been absorbed by their closest corner. This could indicate that transitions between the two intermediate states have become so rapid that the states cease to exist for a comparable length of time compared to the 4 "major" states, while at 273 K they are stable enough to have a comparable timeframe.

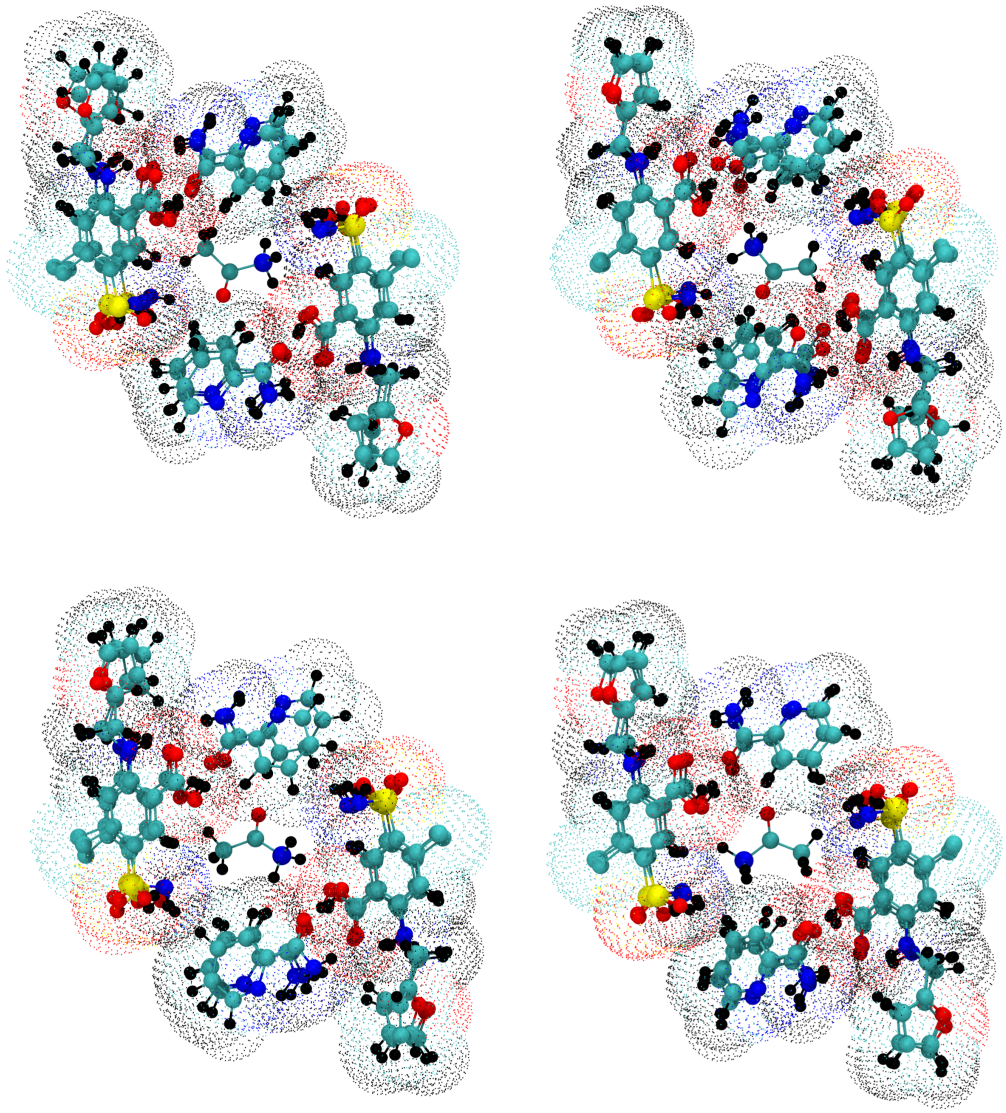

FIGURE 4.21: FSPA-acetone example orientations. The same carbon has been highlighted for each acetone molecule, to highlight the change in C-C vector as well as the change in C-O vector between states. Example orientations for the intermediate states showed a range of orientations, as they can explore a much larger portion of the state space than the "corner" states.

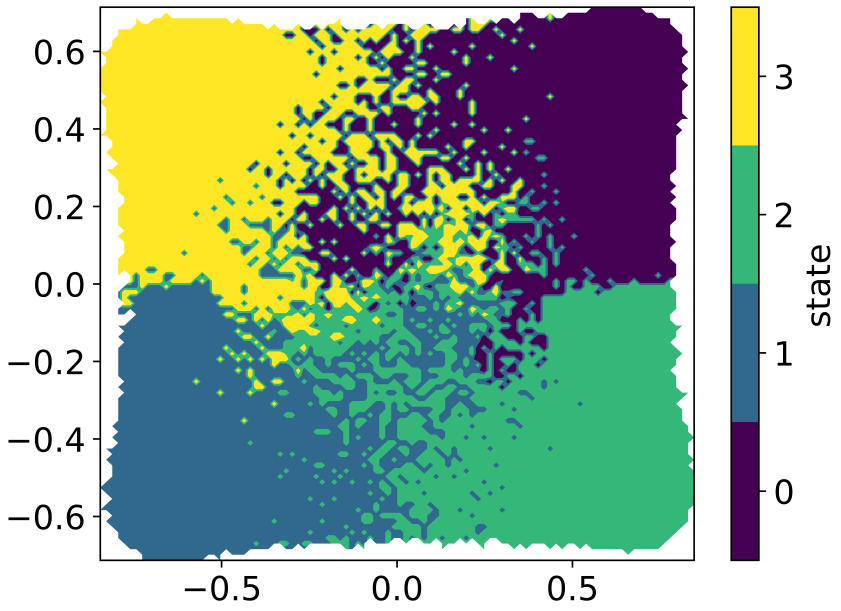

FIGURE 4.22: FSPA-acetone state map for 323 K. This graph shows the division of the free energy diagram projected along TICs 1 and 2 into the different metastable states.

Temperature	Corner->Corner	Side->Side	Top->Bottom	Middle->Corner	Corner->Middle
250	15.47	16.04	11.34	16.77	17.12
263	16.94	17.38	11.27	22.9	25.08
273	4.62	4.75	3.51	4.43	3.33
283	1.9	1.99	1.53	-	-
294	1.62	1.69	1.3	-	-
303	1.17	1.21	0.98	-	-
313	0.99	1.03	0.85	-	-
323	0.7	0.73	0.63	-	-

TABLE 4.4: A table showing the average transition time in ns between the six or four states for the FSPA-ace system for each simulated temperature. Transition times have been grouped according to the type of transition, with average times for each set of states reported. The headings indicate moving between the states as detailed in figure 4.19, with the "corner to corner" transitions indicating transitions between states 2-3, and 4-5, "side to side" between states 3-4 and 2-5 and "top and bottom" transitions between states 5-3 and 2-4.

Table 4.4 shows the average transition times for each type of transition according to the state map (figure 4.19) for the simulations from 250 K to 323 K. From 273 K to 323 K, there is a steady decrease in the transition times, while there is a rather large jump of about 10 ns between 263 and 250. Of interest to note is that the transition between the "bottom" and "top" states is generally the fastest. From the state map, this refers to rotating the acetone molecule about the C-C vector as shown in figure 4.5. Modelling the transition rates using an Arrhenius type expression with the equation: $\tau_c = \tau_0 \exp \frac{E_a}{RT}$ yields activation energies of 24.85 ± 0.6, 24.756 ± 0.6 and 22.59 ± 0.55 kJ mol^{-1} for the corner-corner, side-side and top-bottom transitions respectively. All of these transitions are on the order of a hydrogen bond strength. While the breaking of a hydrogen bond would make sense for the corner-corner and top-bottom transitions, as these reorient the C=O bond, this does not explain the side-side transition also showing this energy barrier.

Analysis of the correlation coefficients showed that for every temperature above 250 K, there was no significant correlation between molecules, with most pairs of molecules having correlation coefficients below 0.5. However, at 250 K, there is an abundance of molecule pairs with coefficients greater than 0.5. This could indicate that at high temperatures, the molecules move randomly flipping between the various states with no interaction from neighbouring molecules, but at lower temperatures, the motions start to become correlated, with one molecule's rotation influencing neighbouring molecules to move. This could indicate co-ordinated motion within the crystal structure of the FSPA molecules. Rotations of the acetone molecules within one channel may distort or require a distortion within the crystal to allow the molecule to move, which could in turn cause a similar distortion further down the crystal, allowing another molecule to also flip.

^{2}H NMR

The data derived from the MD simulations gives us insight into the ^{2}H NMR. Flip rate models and simulated spectra had been generated assuming a single 180° rotation, the rotation axis along the C=O bond. The simulations have shown that two distinct rotations are occuring, often in concert with each other. Using this updated 4-state model gives us transitions that cannot be seen by NMR. When both rotations occur, the average C-D bond vector is reversed, and so the ^{2}H quadrupolar tensor has an equivalent orientation. As a result, the combined four state motion cannot be distinguished from the two state model. The flip rates determined from the analysis are extremely fast, on the order of several nanoseconds for 273 K and less than 2 ns for temperature 283 K and over. As movements on these timescales correspond to motional frequences of several hundred to several thousand MHz, these are well over limit of fast motion, and therefore should not interact with the deuterium quadrupolar coupling constant during measurements.

4.6 Conclusion

In this chapter, we anlaysed the dynamics within a solid pharmaceutical system, focusing on the acetone and ethanol solvates. Using the results of the NMR experiments as a starting point, we applied the method from the previous chapter to observe and model the dynamics present, combining MD simulations with Markov modelling to obtain results that can explain the NMR data and allow insight into the motion of the solvents. The acetone motion has been shown to be more complex than modelled with the NMR, with 4 major states available to it at higher temperatures, and intermediate states available at lower temperatures. These states have been characterised and related to each other through two separate C_2 rotations, with activation

energies on the order of strength of a hydrogen bond. The use of intermolecular vectors has allowed us to gain this insight, as well as reducing the time taken to analyse the data.

The same method was utilised to analyse the ethanol solvate as well, producing a 2 state model, although the TICA free energy diagrams indicate that while additional states may be present, transitions between them are too rapid to produce a reasonable model. The energy barrier obtained from the MD is in good agreement with the NMR experiments, showing that this two state flip is a reasonable explanation of the ethanol motion. The motion is also likely to be uncorrelated, with the ethanol molecules able to flip independently of each other.

Furthering these investigations would include additional simulations at lower temperatures. The motion observed is rapid, and so lowering the temperature would allow us to determine how the molecule moves from state to state, via the suggested rotations or through another pathway. This could also be of use specifically with the ethanol solvate: the presence of additional local energy minima on the TICA free energy diagram suggest that there are more accessible states, but that our current simulations do not spend enough time in them to analyse. Lowering the temperature would keep them there longer, and may help to explain the motion further.

In light of the simulations, the NMR data can be analysed further. The model spectra produced in figure 4.3 assumed only a two-site jump model, so enhancing this model with all four sites would improve the analysis. The addition of the pseudo-C_2 rotation makes physical sense, as the solvent cavity is directionless, allowing relatively free rotation. The MD is providing this plausible model that cannot be directly derived from just x-ray diffraction or NMR. Additionally, selectively labelling one carbon in the acetone system could allow the experiments to see these two motions in action, perhaps

allowing measurement of the rates of each using the NMR. For the ethanol solvate, we can see two major states with each state possibly having three minor states associated with them. This complex motion helps to rationalise the NMR data, as the experimental results do not fit well to simple models. Assuming lower temperatue simulations corroborate this, performing NMR experiments at these low temperatures could allow us to see this fine motion, or clearly seperate the two major orientations available to the ethanol.

5 Conclusions and Future work

5.1 Major conclusions

The work presented in this thesis has the overall aim of utilising Markov state modelling techniques in order to better understand and explain the motions observed through the use of other techniques, most prominently NMR spectroscopy. A workflow was envisaged, in which initial NMR studies would act as a guide for the MD simulations, giving an idea of which kinds of motions to expect and their timescales. The MD would then be performed, which can directly show the motions taking place, using the MSM analysis to determine states and the rates of moving from one state to another. The activation energy obtained from the NMR experiments can be used to validate the MD, as well as determine whether both techniques are observing the same motion. The MD and MSM analysis can then be used to search for other motions, by altering the temperature of simulations or changing between the use of PCA or TICA analysis.

For the diamondoids, this idea was shown to have varying levels of success. The technique was particularly useful for diamantane, clearly showing the C_3 rotation, as well as indicating the presence of 90° rotations and providing agreement in the activation energies calculated from the NMR and the MD. For the triamantane and tetramantane systems, the analysis proved trickier, indicating some key parameters that need to be carefully chosen, as well as some of the limitations of the technique. However, for triamantane, useful

motional data was obtained, showing two different possible motions and the rates and energies of each. For tetramantane, the analysis obtained six distinct states, but would require further work to fully characterise the motions.

In the pharmaceutical systems analysed, the motion of the solvent molecules was investigated. Again, preliminary NMR studies indicated the presence of motion for both systems, as well as a theorised 2 site jump for both systems. While the MD corroborated this for the ethanol solvate, for the acetone it indicated the pesence of four distinct sites, as well as intermediate states at the lower temperatures. The activation energies of these rotations were also in agreement with the NMR, and the motions were determined to be uncorrelated to each other.

In general, the techniques have been shown to have reasonable success in describing the motions present in solid systems. Simple motions can be obtained, on both the long and short timescales, requiring only a quick change in technique to differentiate between the two. Additionally, motion that cannot be seen by NMR or XRD can be found, and the technique naturally shows the presence of other motions along the timescales it can see as well, without specifically searching for them. The semi-automation of the analysis is also useful, allowing a chemist with brief training to accurately use the code, allowing anyone to obtain the same results.

5.2 Future work

Future work on these systems would focus on two aspects: For the FSPA systems, lowering the temperature of the ethanol solvate could give us insight into other potential motions, as there is some evidence of six states being available, with rapid motion linking three states together into another metastable set, as seen so far. Lowering the temperature would freeze this

out, and potentially allows us to see these more rapid motions. For the acetone system, lowering the temperature of the systems could give greater insight into the intermediate states. For the diamondoids, the 90° motion seen could be investigated further, through higher temperature simulations or by extending the lower temperature ones to see this motion. For triamantane and tetramantane, the analysis needs to be refined further, to allow more useful information to be seen. Extending the simulation length for tetramantane and reconsidering the vectors used as input data could aid in this, while longer simulations for the triamantane, or expanding the simulation cell can give more data to work with.

The techniques and workflow described in this work can be applied generally to any solid system, provided several starting pieces of information are available: A unit cell, an appropriately generated force field topology, and a general idea of the motion within the system. Once these pieces of information are obtained, the simulation and subsequent analysis can be performed. While the simulation protocol requires specialist knowledge, the analysis can be semi-automated, allowing a chemist to obtain this data rapidly after some small training. This general applicability could help the analysis become routine in studies of motion within solid systems, due to its accuracy and ease of use.

Bibliography

(1) Husic, B. E.; Pande, V. S. *Journal of the American Chemical Society* **2018**, *140*, PMID: 29323881, 2386–2396.

(2) Bowman, G. R.; Bolin, E. R.; Hart, K. M.; Maguire, B. C.; Marqusee, S. *Proceedings of the National Academy of Sciences* **2015**, *112*, 2734–2739.

(3) Zhuang, W.; Cui, R. Z.; Silva, D.-A.; Huang, X. *The Journal of Physical Chemistry B* **2011**, *115*, PMID: 21388153, 5415–5424.

(4) E., W.; Vanden-Eijnden, E. *Journal of Statistical Physics* **2006**, *123*, 503.

(5) Metzner, P.; Schütte, C.; Vanden-Eijnden, E. *Multiscale Modeling & Simulation* **2009**, *7*, 1192–1219.

(6) Vold, R. L.; Hoatson, G. L. *Journal of Magnetic Resonance* **2009**, *198*, 57 –72.

(7) Hospital, A.; Goñi, J. R.; Orozco, M.; Gelpí, J. L. *Advances and applications in bioinformatics and chemistry : AABC* **2015**, *8*, 26604800[pmid], 37–47.

(8) Schyboll, F.; Jaekel, U.; Petruccione, F.; Neeb, H. *Scientific Reports* **2019**, *9*, 14813.

(9) Hoff, B.; Strandberg, E.; Ulrich, A. S.; Tieleman, D. P.; Posten, C. *Biophysical Journal* **2005**, *88*, 1818–1827.

(10) Huber, T.; Rajamoorthi, K.; Kurze, V. F.; Beyer, K.; Brown, M. F. *Journal of the American Chemical Society* **2002**, *124*, PMID: 11782182, 298–309.

(11) Peng, C.; Atilaw, Y.; Wang, J.; Xu, Z.; Poongavanam, V.; Shi, J.; Kihlberg, J.; Zhu, W.; Erdélyi, M. *ACS Omega* **2019**, *4*, 22245–22250.

(12) Vermeer, L. S.; de Groot, B. L.; Réat, V.; Milon, A.; Czaplicki, J. *European Biophysics Journal* **2007**, *36*, 919–931.

(13) Kerr, H. E.; Softley, L. K.; Suresh, K.; Nangia, A.; Hodgkinson, P.; Evans, I. R. *CrystEngComm* **2015**, *17*, 6707–6715.

(14) Harris, R. K. *Analyst* **2006**, *131*, 351–373.

(15) Bērziņš, A.; Hodgkinson, P. *Solid State Nuclear Magnetic Resonance* **2015**, *65*, NMR Crystallography, 12 –20.

(16) Apperley, D. C.; Markwell, A. F.; Frantsuzov, I.; Ilott, A. J.; Harris, R. K.; Hodgkinson, P. *Phys. Chem. Chem. Phys.* **2013**, *15*, 6422–6430.

(17) Abraham, A.; Apperley, D. C.; Byard, S. J.; Ilott, A. J.; Robbins, A. J.; Zorin, V.; Harris, R. K.; Hodgkinson, P. *CrystEngComm* **2016**, *18*, 1054–1063.

(18) Leelananda, S. P.; Lindert, S. *Beilstein Journal of Organic Chemistry* **2016**, *12*, 2694–2718.

(19) Borhani, D. W.; Shaw, D. E. *Journal of Computer-Aided Molecular Design* **2012**, *26*, 15–26.

(20) Gupta, J.; Nunes, C.; Vyas, S.; Jonnalagadda, S. *The Journal of Physical Chemistry B* **2011**, *115*, PMID: 21306175, 2014–2023.

(21) Young, M. A.; Ravishanker, G.; Beveridge, D. L. *Biophysical journal* **1997**, *73*, 9370428[pmid], 2313–2336.

(22) Chodera, J. D.; Mobley, D. L.; Shirts, M. R.; Dixon, R. W.; Branson, K.; Pande, V. S. *Current Opinion in Structural Biology* **2011**, *21*, 150 –160.

(23) Buch, I.; Giorgino, T.; De Fabritiis, G. *Proceedings of the National Academy of Sciences* **2011**, *108*, 10184–10189.

(24) Rexrode, N. R.; Orien, J.; King, M. D. *The Journal of Physical Chemistry A* **2019**, *123*, PMID: 31099570, 6937–6947.

(25) Xiang, T.-X.; Anderson, B. D. *Molecular Pharmaceutics* **2013**, *10*, PMID: 23116319, 102–114.

(26) Hamad, S.; Moon, C.; Catlow, C. R. A.; Hulme, A. T.; Price, S. L. *The Journal of Physical Chemistry B* **2006**, *110*, PMID: 16494346, 3323–3329.

(27) Vogelsberg, C. S.; Garcia-Garibay, M. A. *Chem. Soc. Rev.* **2012**, *41*, 1892–1910.

(28) Watson, M. A.; Cockroft, S. L. *Chem. Soc. Rev.* **2016**, *45*, 6118–6129.

(29) Wang, Q.; Chen, D.; Tian, H. *Science China Chemistry* **2018**, *61*, 1261–1273.

(30) Iwamura, H.; Mislow, K. *Accounts of Chemical Research* **1988**, *21*, 175–182.

(31) Bedard, T. C.; Moore, J. S. *Journal of the American Chemical Society* **1995**, *117*, 10662–10671.

(32) Godinez, C. E.; Zepeda, G.; Mortko, C. J.; Dang, H.; Garcia-Garibay, M. A. *The Journal of Organic Chemistry* **2004**, *69*, PMID: 14987025, 1652–1662.

(33) Feringa, B. L.; van Delden, R. A.; Koumura, N.; Geertsema, E. M. *Chemical Reviews* **2000**, *100*, PMID: 11777421, 1789–1816.

(34) Catalano, L.; Naumov, P. *CrystEngComm* **2018**, *20*, 5872–5883.

(35) Kottas, G. S.; Clarke, L. I.; Horinek, D.; Michl, J. *Chemical Reviews* **2005**, *105*, PMID: 15826014, 1281–1376.

(36) Karlen, S. D.; Garcia-Garibay, M. A. *Chem. Commun.* **2005**, 189–191.

(37) Albunia, A. R.; Gaeta, C.; Neri, P.; Grassi, A.; Milano, G. *The Journal of Physical Chemistry B* **2006**, *110*, PMID: 17004770, 19207–19214.

(38) Dominguez, Z.; Dang, H.; Strouse, M. J.; Garcia-Garibay, M. A. *Journal of the American Chemical Society* **2002**, *124*, PMID: 12083925, 7719–7727.

(39) Jiang, X.; O'Brien, Z. J.; Yang, S.; Lai, L. H.; Buenaflor, J.; Tan, C.; Khan, S.; Houk, K. N.; Garcia-Garibay, M. A. *Journal of the American Chemical Society* **2016**, *138*, PMID: 26973017, 4650–4656.

(40) Dominguez, Z.; Khuong, T.-A. V.; Dang, H.; Sanrame, C. N.; Nuñez, J. E.; Garcia-Garibay, M. A. *Journal of the American Chemical Society* **2003**, *125*, PMID: 12862478, 8827–8837.

(41) Zimmerman, H. E.; Zhu, Z. *Journal of the American Chemical Society* **1994**, *116*, 9757–9758.

(42) Jarowski, P. D.; Houk, K. N.; Garcia-Garibay, M. A. *Journal of the American Chemical Society* **2007**, *129*, PMID: 17315991, 3110–3117.

(43) Ilott, A. J.; Palucha, S.; Batsanov, A. S.; Wilson, M. R.; Hodgkinson, P. *Journal of the American Chemical Society* **2010**, *132*, PMID: 20334377, 5179–5185.

(44) Ilott, A. J.; Palucha, S.; Hodgkinson, P.; Wilson, M. R. *The Journal of Physical Chemistry B* **2013**, *117*, PMID: 24028495, 12286–12295.

(45) Ilott, A. J.; Palucha, S.; Batsanov, A. S.; Harris, K. D. M.; Hodgkinson, P.; Wilson, M. R. *The Journal of Physical Chemistry B* **2011**, *115*, PMID: 21391533, 2791–2800.

(46) Smith, A. A.; Ernst, M.; Riniker, S.; Meier, B. H. *Angewandte Chemie International Edition* **2019**, *58*, 9383–9388.

(47) Ferreira, T. M.; Ollila, O. H. S.; Pigliapochi, R.; Dabkowska, A. P.; Topgaard, D. *The Journal of Chemical Physics* **2015**, *142*, 044905.

(48) Maisuradze, G. G.; Leitner, D. M. *Proteins: Structure, Function, and Bioinformatics* **2007**, *67*, 569–578.

(49) Altis, A.; Otten, M.; Nguyen, P. H.; Hegger, R.; Stock, G. *The Journal of Chemical Physics* **2008**, *128*, 245102.

(50) Jain, A.; Hegger, R.; Stock, G. *The Journal of Physical Chemistry Letters* **2010**, *1*, 2769–2773.

(51) Markov, A. A. *Izvestiia Fiz.-Matem. Obsch. Kazan Univ.* **1906**, *15*, 135–156.

(52) Anderson, D. F.; Kurtz, T. G. In *Design and Analysis of Biomolecular Circuits: Engineering Approaches to Systems and Synthetic Biology*, Koeppl,

H., Setti, G., di Bernardo, M., Densmore, D., Eds.; Springer New York: New York, NY, 2011, pp 3–42.

(53) Schütte, C.; Fischer, A; Huisinga, W; Deuflhard, P *Journal of Computational Physics* **1999**, *151*, 146 –168.

(54) Green, P. J. *Biometrika* **1995**, *82*, 711–732.

(55) Boulougouris, G. C.; Frenkel, D. *Journal of Chemical Theory and Computation* **2005**, *1*, PMID: 26641505, 389–393.

(56) Boomsma, W. et al. *Journal of Computational Chemistry* **2013**, *34*, 1697–1705.

(57) Bratholm, L. A.; Jensen, J. H. *Chem. Sci.* **2017**, *8*, 2061–2072.

(58) Cavalli, A.; Salvatella, X.; Dobson, C. M.; Vendruscolo, M. *Proceedings of the National Academy of Sciences* **2007**, *104*, 9615–9620.

(59) Shen, Y. et al. *Proceedings of the National Academy of Sciences* **2008**, *105*, 4685–4690.

(60) Hulliger, J.; Hesterberg, R. *Journal of Mathematical Chemistry* **2019**, *57*, 1579–1585.

(61) Wang, Q.; Li, H.-X.; Wang, K.-M.; Wang, X.; Xue, Z.; Jia, L.; Du, L.; Zhao, Q.-H. *Chemistry – An Asian Journal* **2018**, *13*, 1415–1418.

(62) Swope, W. C.; Pitera, J. W.; Suits, F. *The Journal of Physical Chemistry B* **2004**, *108*, 6571–6581.

(63) Swope, W. C.; Pitera, J. W.; Suits, F.; Pitman, M.; Eleftheriou, M.; Fitch, B. G.; Germain, R. S.; Rayshubski, A.; Ward, T. J. C.; Zhestkov, Y.; Zhou, R. *The Journal of Physical Chemistry B* **2004**, *108*, 6582–6594.

(64) Chiang, T.-H.; Hsu, D.; Latombe, J.-C. *Bioinformatics* **2010**, *26*, i269–i277.

(65) Weber, J. K.; Jack, R. L.; Pande, V. S. *Journal of the American Chemical Society* **2013**, *135*, PMID: 23540906, 5501–5504.

(66) Huggins, D. J.; Biggin, P. C.; Dämgen, M. A.; Essex, J. W.; Harris, S. A.; Henchman, R. H.; Khalid, S.; Kuzmanic, A.; Laughton, C. A.; Michel,

J.; Mulholland, A. J.; Rosta, E.; Sansom, M. S. P.; van der Kamp, M. W. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2019**, *9*, e1393.

(67) Plattner, N.; Noé, F. *Nature Communications* **2015**, *6*, 7653.

(68) Bittracher, A.; Banisch, R.; Schütte, C. *The Journal of Chemical Physics* **2018**, *149*, 154103.

(69) Klus, S.; Bittracher, A.; Schuster, I.; Schütte, C. *The Journal of Chemical Physics* **2018**, *149*, 244109.

(70) Reuter, B.; Weber, M.; Fackeldey, K.; Röblitz, S.; Garcia, M. E. *Journal of Chemical Theory and Computation* **2018**, *14*, PMID: 29812922, 3579–3594.

(71) Malmstrom, R. D.; Lee, C. T.; Van Wart, A. T.; Amaro, R. E. *Journal of Chemical Theory and Computation* **2014**, *10*, PMID: 25473382, 2648–2657.

(72) Noé, F.; Rosta, E. *The Journal of Chemical Physics* **2019**, *151*, 190401.

(73) Keller, B.; Hünenberger, P.; van Gunsteren, W. F. *Journal of Chemical Theory and Computation* **2011**, *7*, PMID: 26606352, 1032–1044.

(74) Zhang, W.; Schütte, C. *Entropy* **2017**, *19*, DOI: 10.3390/e19070367.

(75) Sarich, M.; Banisch, R.; Hartmann, C.; Schütte, C. *Entropy* **2014**, *16*, 258–286.

(76) Röblitz, S.; Weber, M. *Advances in Data Analysis and Classification* **2013**, *7*, 147–179.

(77) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Dill, K. A. *Multiscale Modeling & Simulation* **2006**, *5*, 1214–1226.

(78) Wu, H.; Paul, F.; Wehmeyer, C.; Noé, F. *Proceedings of the National Academy of Sciences* **2016**, *113*, E3221–E3230.

(79) Hulliger, J.; Burgener, M.; Hesterberg, R.; Sommer, M.; Brahimi, K.; Aboulfadl, H. *IUCrJ* **2017**, *4*, 360–368.

(80) Hulliger, J. *Chemistry – A European Journal* **2002**, *8*, 4578–4586.

(81) Hart, A. G.; Hansen, T. C.; Kuhs, W. F. *Acta Crystallographica Section A* **2018**, *74*, 357–372.

(82) Riechers, P. M.; Varn, D. P.; Crutchfield, J. P. *Acta Crystallographica Section A* **2015**, *71*, 423–443.

(83) Tang, X.; Bevan, M. A.; Grover, M. A. *Mol. Syst. Des. Eng.* **2017**, *2*, 78–88.

(84) Hodgkinson, P *Progress in Nuclear Magnetic Resonance spectroscopy* **2005**, *46*, 197–222.

(85) Kühne, T. D. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2014**, *4*, 391–406.

(86) Curchod, B. F. E.; Martinez, T. J. *Chemical Reviews* **2018**, *118*, 3305–3336.

(87) Tuckerman, M. E. *Journal of Physics: Condensed Matter* **2002**, *14*, R1297–R1355.

(88) Rapaport, D. C., *The Art of Molecular Dynamics Simulation*, 2nd ed.; Cambridge University Press: 2004.

(89) Young, D., *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*; Wiley-Interscience: 2001.

(90) Jensen, F., *Introduction to Computational Chemistry*; John Wiley and Sons: 2007.

(91) Yu, H.; Duan, D.; Liu, H.; Yang, T.; Tian, F.; Bao, K.; Li, D.; Zhao, Z.; Liu, B.; Cui, T. *Scientific Reports* **2016**, *6*, Article, 18918 EP –.

(92) Takaluoma, T. T.; Laasonen, K.; Laitinen, R. S. *Inorganic Chemistry* **2013**, *52*, PMID: 23540510, 4648–4657.

(93) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. *Chemical Reviews* **2016**, *116*, 7898–7936.

(94) Ingolfsson, H. I.; Lopez, C. A.; Uusitalo, J. J.; de Jong, D. H.; Gopal, S. M.; Periole, X.; Marrink, S. J. *Wiley interdisciplinary reviews - Computational molecular science* **2014**, *4*, 225–248.

(95) Sponer, J.; Bussi, G.; Krepl, M.; Banas, P.; Bottaro, S.; Cunha, R. A.; Gil-Ley, A.; Pinamonti, G.; Poblete, S.; Jureacka, P.; Walter, N. G.; Otyepka, M. *Chemical reviews* **2018**, *118*, 4177–4338.

(96) Huang, J.; MacKerell Jr, A. D. *Journal of Computational Chemistry* **2013**, *34*, 2135–2145.

(97) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell Jr., A. D. *Journal of Computational Chemistry* **2010**, *31*, 671–690.

(98) Vanommeslaeghe, K.; MacKerell, A. D. *Journal of Chemical Information and Modeling* **2012**, *52*, PMID: 23146088, 3144–3154.

(99) Verlet, L. *Phys. Rev.* **1967**, *159*, 98–103.

(100) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *The Journal of Chemical Physics* **1984**, *81*, 3684–3690.

(101) Frenkel, D.; Smit, B. In *Understanding Molecular Simulation (Second Edition)*, Frenkel, D., Smit, B., Eds., Second Edition; Academic Press: San Diego, 2002, pp 139 –163.

(102) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695–1697.

(103) Parrinello, M.; Rahman, A. *Journal of Applied Physics* **1981**, *52*, 7182–7190.

(104) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. *Acta Crystallographica Section B* **2016**, *72*, 171–179.

(105) Berendsen, H.; van der Spoel, D.; van Drunen, R. *Computer Physics Communications* **1995**, *91*, 43 –56.

(106) Lindahl, E.; Hess, B.; van der Spoel, D. *Molecular modeling annual* **2001**, *7*, 306–317.

(107) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *Journal of Computational Chemistry* **2005**, *26*, 1701–1718.

(108) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *Journal of Chemical Theory and Computation* **2008**, *4*, PMID: 26620784, 435–447.

(109) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.;

Lindahl, E. *Bioinformatics (Oxford, England)* **2013**, *29*, 23407358[pmid], 845–854.

(110) Páll, S.; Abraham, M. J.; Kutzner, C.; Hess, B.; Lindahl, E. In *Solving Software Challenges for Exascale*, ed. by Markidis, S.; Laure, E., Springer International Publishing: Cham, 2015, pp 3–27.

(111) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. *SoftwareX* **2015**, *1-2*, 19 –25.

(112) Gagniuc, P. A., *Markov Chains: From Theory to Implementation and Experimentation*; John Wiley and Sons: NJ, USA, 2017.

(113) MacQueen, J. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, University of California Press: Berkeley, Calif., 1967, pp 281–297.

(114) Lloyd, S. *IEEE Transactions on Information Theory* **1982**, *28*, 129–137.

(115) Baum, L. E.; Petrie, T.; Soules, G.; Weiss, N. *Ann. Math. Statist.* **1970**, *41*, 164–171.

(116) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. *Journal of Chemical Theory and Computation* **2015**, *11*, 5525–5542.

(117) Metzner, P.; Schütte, C.; Vanden-Eijnden, E. *Multiscale Modeling & Simulation* **2008**, *7*, 1192–1219.

(118) Berezhkovskii, A.; Hummer, G.; Szabo, A. *The Journal of Chemical Physics* **2009**, *130*, 205102.

(119) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proceedings of the National Academy of Sciences* **2009**, *106*, 19011–19016.

(120) Michils, A. *Bulletin des Sociétés Chimiques Belges* **1948**, *57*, 575–617.

(121) Timmermans, J *Journal of Physics and Chemistry of Solids* **1961**, *18*, 1–8.

(122) Gunawan, M. A.; Hierso, J.-C.; Poinsot, D.; Fokin, A. A.; Fokina, N. A.; Tkachenko, B. A.; Schreiner, P. R. *New J. Chem.* **2014**, *38*, 28–41.

(123) Gunawan, M. A.; Poinsot, D.; Domenichini, B.; Dirand, C.; Chevalier, S.; Fokin, A. A.; Schreiner, P. R.; Hierso, J.-C. *Nanoscale* **2015**, *7*, 1956–1962.

(124) Amoureux, J. P.; Bee, M.; Damien, J. C. *Acta Crystallographica Section B* **1980**, *36*, 2633–2636.

(125) Nitzan, A.; Ratner, M. A. *Science* **2003**, *300*, 1384–1389.

(126) Schnurpfeil, A.; Albrecht, M. *Theoretical Chemistry Accounts* **2007**, *117*, 29–39.

(127) Sinkel, C.; Agarwal, S.; Fokina, N. A.; Schreiner, P. R. *Journal of Applied Polymer Science* **2009**, *114*, 2109–2115.

(128) Ghosh, A.; Sciamanna, S. F.; Dahl, J. E.; Liu, S.; Carlson, R. M. K.; Schiraldi, D. A. *Journal of Polymer Science Part B: Polymer Physics* **2007**, *45*, 1077–1089.

(129) Sciamanna, S.; Liu, S.; Mukai, A. Diamondoid-based nucleating agents for thermoplastics, US20070037909A1.

(130) Kabo, G.; Blokhin, A.; Charapennikau, M.; Kabo, A.; Sevruk, V. *Thermochimica Acta* **2000**, *345*, 125 –133.

(131) Britcher, A. R.; Strange, J. H. *J. Chem. Soc., Faraday Trans. 2* **1978**, *74*, 1767–1777.

(132) Angell, C. *Journal of Non-Crystalline Solids* **1991**, *131-133*, Proceedings of the International Discussion Meeting on Relaxations in Complex Systems, 13 –31.

(133) Astumian, R. D.; Mukherjee, S.; Warshel, A. *Chemphyschem : a European journal of chemical physics and physical chemistry* **2016**, *17*, 27149926[pmid], 1719–1741.

(134) Williams, V. Z.; von Ragué Schleyer, P.; Gleicher, G. J.; Rodewald, L. B. *Journal of the American Chemical Society* **1966**, *88*, 3862–3863.

(135) Carrell, H.; Donohue, J. *Tetrahedron Letters* **1969**, *10*, 3503 –3504.

(136) Cernik, R.; Evans, E.; Hine, R.; Richards, J. *Solid State Communications* **1978**, *27*, 1017 –1019.

(137) Wickins, H. M. *Master's thesis* **2017**.

(138) Apperley, D. C.; Harris, R. K.; Hodgkinson, P., *Solid-state NMR : basic principles & practice*; Momentum Press: [New York, N.Y.] (222 East 46th Street, New York, NY 10017), 2012.

(139) Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D. *Journal of Chemical Information and Modeling* **2012**, *52*, PMID: 23145473, 3155–3168.

(140) Bowman, G. R.; Pande, V. S.; Noé, F., *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*; Springer Netherlands: 2014.

(141) Chatterjee, A.; Voter, A. F. *The Journal of Chemical Physics* **2010**, *132*, 194101.

(142) https://www.drugbank.ca/drugs/DB00695 Furosemide, accessed January 2020.

(143) Goud, N. R.; Gangavaram, S.; Suresh, K.; Pal, S.; Manjunatha, S. G.; Nambiar, S.; Nangia, A. *Journal of Pharmaceutical Sciences* **2012**, *101*, 664–680.

(144) Granero, G.; Longhi, M.; Mora, M.; Junginger, H.; Midha, K.; Shah, V.; Stavchansky, S.; Dressman, J.; Barends, D. *Journal of Pharmaceutical Sciences* **2010**, *99*, 2544 –2556.

(145) Wickins, H. M. *Unpublished FSPA results* **2020**.

(146) Hannah, E. K. NMR Crystallography of Disordered Cocrystals, Ph.D. Thesis, Durham University, 2017.

(147) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. *Journal of Chemical Theory and Computation* **2015**, *11*, PMID: 26574453, 3696–3713.

(148) Ponder, J. W.; Case, D. A. In *Protein Simulations*; Advances in Protein Chemistry, Vol. 66; Academic Press: 2003, pp 27 –85.

(149) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *Journal of Computational Chemistry* **2004**, *25*, 1157–1174.

(150) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *Journal of Molecular Graphics and Modelling* **2006**, *25*, 247 –260.

(151) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. *Journal of Computational Physics* **1977**, *23*, 327 –341.

(152) Yoneya, M.; Berendsen, H. J. C.; Hirasawa, K. *Molecular Simulation* **1994**, *13*, 395–405.

A Appendix A: Data Tables

A.1 Chapter 3: Diamondoids

State from\State to	0	1	2
0	-	4.91	4.70
1	4.97	-	4.71
2	4.90	4.86	-

TABLE A.1: The transition rates from state to state in ns for the 310 K system.

State from\State to	0	1	2
0	-	3.04	2.90
1	3.06	-	2.89
2	3.04	3.02	-

TABLE A.2: The transition rates from state to state in ns for the 325 K system.

State from\State to	0	1	2
0	-	1.33	1.35
1	1.35	-	1.35
2	1.37	1.35	-

TABLE A.3: The transition rates from state to state in ns for the 350 K system.

Residue 1	Residue 2	Correlation coefficient value
17	17	0.999
23	23	0.999
44	44	0.999
60	60	0.999
101	101	0.999

TABLE A.4: An excerpt of the correlation coefficients for the 350 K system. Coefficient values can range from 0 for entirely uncorrelated to 1/-1 for entirely postiively/negatively correlated.

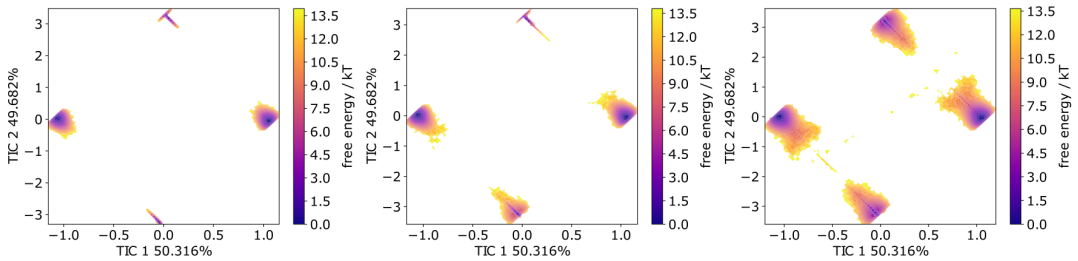

FIGURE A.1: The TIC 1/2 plot for triamantane. From left to right, the temperatures are 310 K, 350 K and 400 K.

State from\State to	0	1	2	3
0	-	16.27±4.85	1.16±0.62	8.51±2.88
1	56.55±51.56	-	20.62±15.69	2.40±2.11
2	35.93±39.27	15.11±4.7	-	7.35±2.61
3	54.15±51.14	7.76±2.95	18.22±14.99	-

TABLE A.5: The full transition times from state to state in μs for the 400 K system. The high uncertainties on many of the values is indicative of the slow speed of these transitions, and the lack of transition events present in the simulation. The simulation lasted 400 ns, and yet the model is predicting transition times of at least one order of magnitude above this.

A.2 Chapter 4: FSPA

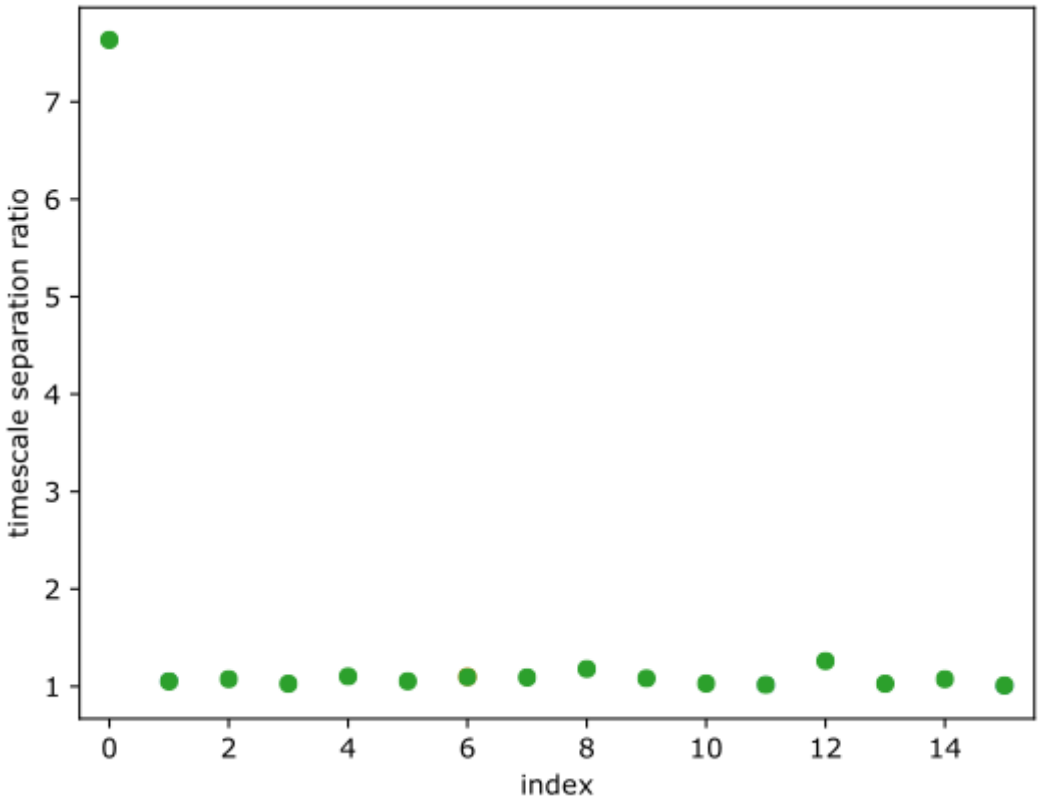

FIGURE A.2: FSPA-ethanol timescale separation plot. Each point shows the ratio of timescale n and timescale $n+1$, with index 0 showing the ratio between timescales 1 and 2.